A model-based method for 3D reconstruction of cerebellar parallel fibres from high-resolution electron microscope images

Martin T. O'Reilly

ORCID: 0000-0002-1191-3492

A thesis submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London

UCL Centre for Mathematics and Physics in the Life Sciences and Experimental Biology

UCL Department of Computer Science

UCL Wolfson Institute for Biomedical Research

Declaration

I, Martin O'Reilly, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signed: *Date:*

Licence

"If you want to understand function, study structure"

- Francis Crick, What a mad pursuit.

Abstract

In order to understand how the brain works, we need to understand how its neural circuits process information. Electron microscopy remains the only imaging technique capable of providing sufficient resolution to reconstruct the dense connectivity between all neurons in a circuit. Automated electron microscopy techniques are approaching the point where usefully large circuits might be successfully imaged, but the development of automated reconstruction techniques lags far behind. No fully-automated reconstruction technique currently produces acceptably accurate reconstructions, and semi-automated approaches currently require an extreme amount of manual effort. This reconstruction bottleneck places severe limits on the size of neural circuits that can be reconstructed. Improved automated reconstruction techniques are therefore highly desired and under active development. The human brain contains $\sim$86 billion neurons and $\sim$80% of these are located in the *cerebellum*. Of these cerebellar neurons, the vast majority are *granule cells*. The axons of these granule cells are called *parallel fibres* and tend to be oriented in approximately the same direction, making 2+1D reconstruction approaches feasible. In this work we focus on the problem of reconstructing these *parallel fibres* and make four main contributions: (1) a model-based algorithm for reconstructing 2D parallel fibre cross-sections that achieves state of the art 2D reconstruction performance; (2) a fully-automated algorithm for reconstructing 3D parallel fibres that achieves state of the art 3D reconstruction performance; (3) a semi-automated approach for reconstructing 3D parallel fibres that significantly improves reconstruction accuracy compared to our fully-automated approach while requiring $\sim$40$\times$ less labelling effort than a purely manual reconstruction; (4) a "gold standard" ground truth data set for the molecular layer of the mouse cerebellum that will provide a valuable reference for the development and benchmarking of reconstruction algorithms.

Acknowledgements

My family and friends have provided much appreciated support, especially during the thesis writing process. In particular, the proof-reading assistance provided by Peter O'Reilly was invaluable. My supervisors Lewis Griffin and Arnd Roth have provided academic support and encouragement throughout my PhD. Michael Häusser has also provided significant support for this work. In addition to providing the electron microscope and the team that performed the ground truth labelling, the Häusser lab has provided a stimulating scientific environment. This work was supported by an EPSRC PhD stipend from the UCL CoMPLEX Doctoral Training Centre, as well as additional financial support from Michael Häusser at the UCL Wolfson Institute for Biomedical Research and Lewis Griffin at the UCL Department of Computer Science.

The manual ground truth labelling on which this work completely relies would not have been possible without the combined efforts of a large team from the Häusser Lab at the UCL Wolfson Institute for Biomedical Research. Sample preparation and imaging were performed by Sarah Rieubland, Arnd Roth and Arifa Naeem. The manual labelling of these images was performed by Sophie Gordon-Smith, Rashmi Gamage, Trisha Patel, Kylie Wong and Maja Boznakova. Both Maja Boznakova and Sarah Rieubland were a great help in managing the labelling process, and Arnd Roth provided valuable support throughout the process.

Contents

1 Introduction **14**

1.1 Motivation and approach . . . 14

1.2 Problem statement . . . 16

1.3 Contributions of this work . . . 16

1.4 Thesis outline . . . 18

1.5 Publications . . . 19

2 Reconstructing the connectome **20**

2.1 Connectomics . . . 20

2.1.1 What is connectomics? . . . 20

2.1.2 Understanding neural connectivity . . . 21

2.1.3 Methods for connectomics . . . 24

2.1.4 Scale of the challenge . . . 26

2.1.5 Focus on cerebellar parallel fibre reconstruction . . . 28

2.2 Imaging for connectomics . . . 29

2.2.1 Light microscopy . . . 29

2.2.2 Electron microscopy . . . 31

3 Image segmentation for connectomics **35**

3.1 Pixel based approaches . . . 35

3.1.1 Techniques for learning to predict pixels . . . 36

3.1.2 Image features . . . 38

3.1.3 Intracellular structure identification . . . 39

3.1.4 Pixel grouping . . . 40

3.1.5 Linking cross-sections across slices . . . 40

3.2 Contour-based approaches . . . 41

3.3 Issues with pixel and contour based approaches . . . 42

3.4 Semi-automated approaches . . . 43

3.5 Measures of segmentation accuracy . . . 45

3.5.1 Binary pixel classification accuracy . . . 45
3.5.2 High level similarity measures . . . 50
3.5.3 Overlap as a measure of similarity . . . 51
3.5.4 Measures of 3D reconstruction accuracy . . . 53

4 Data collection and curation **54**
4.1 Image acquisition . . . 54
4.2 Generating ground truth labels . . . 54
4.2.1 Initial 2D labelling . . . 54
4.2.2 Full 2D membrane labelling . . . 55
4.2.3 3D neurite labelling . . . 55
4.2.4 Data quality . . . 56
4.3 Publication of a "gold standard" reference data set . . . 58

5 Modelling neural fibres **60**
5.1 Overview . . . 60
5.2 Circles as a representation of fibre cross-sections . . . 61
5.2.1 Circles have high overlap with fibre cross-sections . . . 61
5.2.2 Circles are a suitable abstraction for modelling neural fibres . . . 63
5.3 Finding representative circles from overlap . . . 70
5.3.1 Ground truth overlap as a "fibreness" score . . . 70
5.3.2 Finding circles from ground truth overlap . . . 71
5.3.3 Predicting overlap is sufficient for finding good quality circles . . . 71
5.4 Predicting overlap from image features . . . 72
5.4.1 Basic Image Features (BIFs) . . . 72
5.4.2 Annular BIF histograms as "fibreness" feature vectors . . . 77

6 Reconstructing fibre cross-sections **80**
6.1 Algorithm overview . . . 80
6.1.1 Learning a mapping from EM image to ground truth overlap . . . 80
6.1.2 Finding fibre circles . . . 83
6.2 Training data selection . . . 83
6.2.1 Range of circle radii . . . 83
6.2.2 Picking training circles . . . 85
6.3 Selecting a "fibreness" feature vector . . . 86

6.3.1 BIF and annulus parameters . . . 86
6.3.2 BIF type and histogram normalisation . . . 88
6.3.3 Adding non-BIF image features . . . 91
6.4 Selecting a method to predict ground truth overlap . . . 91
6.4.1 Linear regression . . . 92
6.4.2 Logistic regression . . . 93
6.4.3 Random forest regression . . . 93
6.4.4 Comparison of regression methods . . . 94
6.4.5 Combining classification and regression . . . 96
6.5 Algorithm performance . . . 96
6.5.1 Visualising algorithm performance . . . 96
6.5.2 Analysing algorithm performance . . . 97
6.5.3 Properties of poorly found fibres . . . 99
6.6 Comparing reconstruction methods across studies . . . 101
6.6.1 Issues when comparing results across studies . . . 101
6.6.2 Existing benchmarks and comparable results . . . 104
6.7 Benchmarking against ilastik on our data set . . . 106
6.7.1 The ilastik reconstruction pipeline . . . 106
6.7.2 Optimising the ilastik reconstruction . . . 107
6.7.3 Making a fair comparison . . . 108
6.7.4 Ilastik reconstruction accuracy . . . 110
6.7.5 Comparison of reconstruction accuracy . . . 111
6.8 Future work . . . 111
6.8.1 3D features . . . 111
6.8.2 Combining algorithms . . . 113

7 Reconstructing fibres in three dimensions 114
7.1 Algorithm overview . . . 114
7.2 Evaluating 3D reconstruction accuracy . . . 116
7.2.1 Matched f-measure . . . 116
7.2.2 Matched segment run-length . . . 117
7.2.3 Previously reported measures . . . 117
7.2.4 Mitigating edge effects . . . 120
7.3 Selecting tube finding parameters . . . 121

7.3.1 Initial parameter sensitivity exploration . . . 121
7.3.2 Balancing matched segment run-length and matched f-measure . . . 123
7.3.3 Visualising individual reconstructed fibres . . . 128
7.3.4 Summarising 3D reconstruction accuracy . . . 130
7.3.5 Permitting temporary tracking failures . . . 130
7.3.6 Accounting for censored data . . . 133
7.4 Benchmarking against another mouse cerebellum study . . . 135
7.4.1 Image z-resolution . . . 136
7.4.2 Level of tracking failure permitted . . . 136
7.4.3 Reported run-length measure . . . 136
7.4.4 Summary of benchmark results . . . 141
7.5 A semi-automated approach . . . 142
7.5.1 Combining manual and algorithmic inputs . . . 142
7.5.2 The spatial influence of manual labelling . . . 143
7.5.3 The effect of labelling effort on reconstruction accuracy . . . 145
7.6 Limitations and further work . . . 149
7.6.1 Limitations of our work . . . 149
7.6.2 Further work . . . 152

8 Conclusions **153**
8.1 Contributions . . . 153
8.1.1 A model-based algorithm for 2D reconstruction of fibre cross-sections . . . 153
8.1.2 A fully-automated algorithm for reconstructing 3D parallel fibres . . . 154
8.1.3 A semi-automated approach for reconstructing 3D parallel fibres . . . 155
8.1.4 A “gold standard” ground truth for the mouse cerebellum . . . 155
8.2 Issues and limitations . . . 155
8.2.1 Restriction to 2+1D problems . . . 155
8.2.2 Benchmarking difficulties . . . 156
8.3 Future work . . . 156
8.3.1 Refined 3D benchmarking . . . 156
8.3.2 A neuroscientific analysis of our ground truth reconstruction . . . 156
8.3.3 Tube finding in three dimensions . . . 156
8.3.4 Combining different reconstruction methods . . . 157

Appendices **157**

A Sample preparation and imaging **158**

A.1 Sample preparation 158

A.2 Image acquisition 159

A.3 Image post-processing 159

B Jaccard index publication history **160**

C Generating BIFs **162**

Bibliography **162**

List of Figures

2.1 From electron microscope images to cerebellar model 30
2.2 EM image of mouse cerebellum 34

3.1 Binary confusion matrix 46
3.2 Definition of overlap 52

4.1 Ground truth labels 57

5.1 Circle-based overlap and threading 62
5.2 Distribution of overlap between ground truth polygons and circles 64
5.3 Example fibre polygons and maximum overlap circles 65
5.4 Synapse attribution 67
5.5 Ground truth overlap volume 70
5.6 Overlap distribution for ground truth overlap volume circles 73
5.7 Fibre circles found using the ground truth overlap volume 74
5.8 EM image and BIFs for a region of mouse cerebellum 77
5.9 Annular BIF histogram 78
5.10 Radially normalised BIFs (rBIFs) 79

6.1 Overview of 2D fibre finding algorithm 81
6.2 Distribution of radii for ground truth fibre circles 85
6.3 BIF histogram parameter sensitivity 89
6.4 Visualising the effect of BIF parameters 90
6.5 BIF type and normalisation 92
6.6 Regression method 95
6.7 Combining classification and regression 97
6.8 Fibre circles found using the best algorithm parameters 98
6.9 Overlap distribution for circles found by algorithm 100
6.10 Overlap of found fibres as a function of true fibre properties 101
6.11 2D segmentation benchmark comparison 110
6.12 Cell segments found using the best ilastik parameters 112

7.1 Calculation of various 3D segmentation measures 120
7.2 Tube finding parameter sensitivity with no minimum found fibre length 122
7.3 Effect of enforcing a minimum fibre length 123
7.4 Run-length vs. matched f-measure trade-off 125
7.5 Matched segments for found fibres 126
7.6 Matched segments for true fibres 127
7.7 Tube finding parameter sensitivity with an enforced minimum fibre length . . . 129
7.8 Example fully-automated found fibres 131
7.9 Run-length distribution for a selection of parameter sets 132
7.10 Permitting temporary tracking failures 133
7.11 Accounting for censored data 134
7.12 Validating Kaplan-Meier estimation 135
7.13 Benchmark comparison 139
7.14 Spatial influence of manual labelling 144
7.15 Maximising run-length with a semi-automated approach 146
7.16 Maximising f-measure with a semi-automated approach 148
7.17 Example semi-automated found fibres 150

List of Tables

2.1 Connectome reconstruction effort 27

3.1 A family of binary similarity measures 48
3.2 Precision and recall based similarity measures 49

4.1 Inter-person labelling variability 59

5.1 BIF classes 75
5.2 Derivative of Gaussian filters 76

7.1 Estimating the f-measure for the Jurrus benchmark data 140

Chapter 1

Introduction

1.1 Motivation and approach

In order to understand how the brain works, we need to understand how its neural circuits process information. While functional recording techniques are improving in the number and density of neurons they can simultaneously record from, they are not currently capable of reconstructing the functional connectivity between all neurons comprising a processing circuit. Electron microscopy remains the only technique capable of reconstructing the dense connectivity between all neurons in a circuit.

Automated reconstruction techniques

The reconstruction of connectivity from electron microscope images currently requires an extreme amount of manual effort. The first *connectome* describing the connectivity between all 302 neurons in the nematode worm *C.elegans* took 10-15 years to complete in the 1970s and 80s. However, despite advances in automated imaging and reconstruction techniques, this remains the only whole-animal connectome in existence to this day. Automated imaging techniques are approaching the point where an entire 1 mm^3 cortical column might be successfully imaged. However, the development of automated reconstruction techniques lags far behind. No fully-automated method for reconstructing neurons from electron microscope images currently produces acceptably accurate reconstructions without substantial human proof-reading and correction. Even the most efficient semi-automated approaches currently available would require 140 years of manual labelling effort to reconstruct a 1 mm^3 cortical column. This rises to $\sim$70,000 years for an entire mouse brain. Improved automated reconstruction techniques are therefore highly desired and under active development.

Issues with existing approaches

Current automated reconstruction techniques fall into one of two main categories. Bottom-up *pixel-based* approaches predict the probability that each pixel is *membrane* or *non-membrane*

based on relatively local image features. They then group pixels into clusters of *non-membrane* pixels that are sufficiently well separated by *membrane* pixels. While a range of sophisticated techniques can be applied to the problem of generating a segmentation from such membrane probability maps, these techniques often suffer from the restricted spatial context considered when classifying pixels. This can result in the misclassification of pixels where membrane evidence is locally misleading, and small local errors in pixel classification can result in significant topological errors in segmentation. While several studies have explored methods to increase this context, none have fully solved the problem. In contrast, top-down *contour-based* methods model neurite boundaries as closed contours and fit these directly to the image data, often including both geometrical constraints (e.g. convexity) and interaction constraints (e.g. limiting overlap between contours). Therefore locally weak membrane evidence can be "bridged" by a contour if the evidence along the rest of the perimeter is sufficiently strong. However, the solution space of possible sets of contours is too vast to exhaustively evaluate and can only be searched by local refinement. It is likely to contain many local optima and so the quality of the found "optimal" solution is highly dependent on where the search starts. Thus, while this method is effective for propagating a known good set of contours to an adjacent slice, it is much less effective when a good quality initialisation is unavailable.

A novel model-based approach

We propose an alternative model-based approach that considers a larger spatial context than pixel-based methods while having a much more restricted solution space than contour-based methods. We model the cross-sections of neurites as circles, which addresses the key issues with both pixel-based and contour-based approaches. We evaluate the image evidence for each circle within an annular region around its perimeter. This results in the consideration of evidence from a larger context than most pixel-based methods, and permits us to integrate image evidence over the entire boundary of a fibre cross-section in a similar manner to contour-based methods. The use of circles as our model of fibre cross-sections results in a drastic reduction in the number of degrees of freedom compared to contour-based methods. This permits us to evaluate the evidence provided by the image for a full range of candidate circles at each pixel. This exhaustive evaluation of the solution space avoids the problem of local minima associated with contour-based methods.

Focus on cerebellar parallel fibres

The human brain contains $\sim$86 billion neurons, and $\sim$80% of these are located in the *cerebellum* (Azevedo et al., 2009). This fist-sized region at the back of the brain is crucial for the co-ordination of motion and the learning of new motor programs. Therefore, understanding

how information is processed in the cerebellum is of great interest. The vast majority of the ∼69 billion neurons in the cerebellum are *granule cells*, which form the input layer of the cerebellum and receive a range of motor and sensory inputs. The axons of these cells are known as *parallel fibres*, and provide the primary input to the *Purkinje cells* that provide the sole output from the cerebellum. Accurately reconstructing these *parallel fibres* is therefore crucial in order to understand the circuitry of the cerebellum. Their vast number, long length and small diameter make this a challenging problem. However, their parallel nature means that imaging can be performed such that most fibres run approximately perpendicular to the image plane, lending themselves to a less computationally intensive 2+1D reconstruction approach. In such an approach, fibre cross-sections are identified in each image independently, and then combined into 3D fibres in a separate step. Several studies have taken such a 2+1D approach to the reconstruction of neural fibres, and this is the approach we take in this work.

1.2 Problem statement

We address the problem of reconstructing 3D parallel fibres from classically stained electron microscope images of the molecular layer of the mouse cerebellum. We take a 2+1D approach, first reconstructing 2D fibre cross-sections independently in each image slice and then linking these cross-sections together across slices to form 3D fibre reconstructions. Our data set also includes segments of other neurites such as interneuron axons, interneuron dendrites, Purkinje cell dendrites, glial cells and some possible climbing fibres. In this work we focus on the reconstruction of *parallel fibre* axons, but also attempt to reconstruct any other axons that run within ∼45° of the image plane. Collectively we refer to such axons as *fibres*. Reconstructing other neurites present in the data set is outside the scope of this work, as is identifying the synapses that form the connections between neurites. Our approach is specific to this region of the brain, where a large proportion of fibres are oriented approximately parallel to each other, and we do not claim our approach is suitable for other brain regions with less regular structure.

1.3 Contributions of this work

In this work we make four main contributions..

1. A model-based algorithm for reconstructing 2D parallel fibre cross-sections that achieves state of the art 2D reconstruction performance.

2. A fully-automated algorithm for reconstructing 3D parallel fibres that achieves state of the art 3D reconstruction performance.

3. A semi-automated approach for reconstructing 3D parallel fibres that significantly improves reconstruction accuracy compared to our fully-automated approach while requiring $\sim$40$\times$ less labelling effort than a purely manual reconstruction.

4. A "gold standard" ground truth data set for the molecular layer of the mouse cerebellum that will provide a valuable reference for the development and benchmarking of reconstruction algorithms.

A model-based algorithm for reconstructing 2D parallel fibre cross-sections

We develop a model-based algorithm for the reconstruction of 2D parallel fibre cross-sections in classically stained electron microscopy images of the cerebellum. We benchmark our algorithm against *ilastik*, a state of the art pixel-based algorithm (section 6.7). The performance of our algorithm and *ilastik* are very similar, achieving $\sim$50% on an overlap-based f-measure. We would therefore claim state of the art performance at reconstructing 2D parallel fibre cross-sections. Our approach extends the restricted spatial context associated with bottom-up *pixel-based* methods, while avoiding the unmanageably large solution spaces associated with top-down *countour-based* methods. To achieve this we model fibre cross-sections as circles (chapter 5).

A fully-automated algorithm for reconstructing 3D parallel fibres

We develop a fully-automated algorithm for combining the 2D cross-sections generated by our model-based algorithm into 3D tubes representing neural fibres. We benchmark our algorithm against 3D results recently reported by another group on similar mouse cerebellum data (section 7.4). While there are some difficulties making an accurate cross-study comparison, our algorithm appears to comfortably outperform this benchmark. We would therefore claim state of the art performance at reconstructing 3D parallel fibres.

A semi-automated approach for reconstructing 3D parallel fibres

While our fully-automated algorithm achieves state of the art performance for reconstructing parallel fibres in 3D, it falls far short of the accuracy required to generate fully-automated reconstructions of neural circuits of an interesting size. We therefore develop a semi-automated approach which combines sparse 2D manual labelling with our 3D reconstruction algorithm (section 7.5.1). This results in significant improvements to the reconstruction accuracy achieved compared to our fully-automated algorithm, while achieving a reduction in labelling effort of $\sim$40$\times$ compared to a purely manual reconstruction. However, additional proof-reading of our semi-automated reconstruction is still necessary to correct the remaining errors. We have yet to quantify the additional manual effort required for these corrections, and this will reduce our

final achieved efficiency gain. However, we would expect our fully-corrected semi-automated approach to remain significantly more efficient than a purely manual approach.

A "gold standard" ground truth for the molecular layer of the mouse cerebellum

We have manually labelled all extracellular membrane and neurite interiors in a 23.7×7.9×4.6 μm region of the molecular layer of the mouse cerebellum (2548×852×512 pixels). Each neurite has been given a consistent 3D label across all its cross-sections, generating a true 3D ground truth. We are currently using the manually reconstructed ground truth data to analyse the ultrastructure of the molecular layer of the cerebellum. Once our ultrastructure analysis is published, we will publish both the electron microscope images and the ground truth labelling for this data set in the open access Cell Centered Database (CCDB). This will provide a valuable reference data set for the development and benchmarking of current and future reconstruction algorithms. It will also help accelerate the expansion of the field to include researchers without access to neuroscience collaborators and electron microscopes.

1.4 Thesis outline

In **chapter 2** we review the concept of a *connectome*, which describes the connectivity between all neurons in an organism or brain region. We describe the uses of a connectome and the challenges of generating one. We then discuss the selection of the cerebellum as our brain region of interest, and the various imaging techniques that have been brought to bear on the problem of reconstructing connectomes. In **chapter 3** we discuss the range of image segmentation methods that have been applied to the problem of reconstructing connectomes, and briefly explain how our model-based method relates to these. We then discuss a range of segmentation accuracy measures, and explain the selection of *overlap* as the basis for our chosen measure. In **chapter 4** we describe the collection and curation of our image data and ground truth labelling. In **chapter 5** we describe our model-based approach in more detail, justifying our selection of circles to represent parallel fibre cross-sections. We then demonstrate that predicting the *overlap* of a set of candidate circles with the fibre cross-sections within an image is sufficient to generate a high quality reconstruction. In **chapter 6** we describe our algorithm for reconstructing 2D parallel fibre cross-sections. We describe the selection of optimal algorithm parameters, and evaluate the performance of our algorithm by benchmarking it against *ilastik*, a state of the art pixel-based method. The performance of our algorithm and *ilastik* are very similar, achieving ∼50% on an overlap-based f-measure. Although we evaluate both algorithms on our data set, it is difficult to make a direct comparison due to differences in the density of the reconstructions provided by both algorithms. We discuss these issues and justify the va-

lidity of our benchmark comparison. In **chapter 7** we describe our fully- and semi-automated algorithms for reconstructing parallel fibres in 3D. We describe our fully-automated approach, which combines the 2D cross-sections generated by our 2D algorithm into 3D tubes. We describe the selection of optimal algorithm parameters, and benchmark our algorithm against 3D results recently reported by another group on similar mouse cerebellum data. Our algorithm appears to comfortably outperform the benchmark, although there are some difficulties making an accurate cross-study comparison. We discuss these issues and justify the validity of our benchmark comparison. We also describe a semi-automated approach, which combines sparse 2D manual labelling with our 3D reconstruction algorithm. We evaluate the effect of manual labelling effort on reconstruction accuracy, and achieve a significant improvement compared to our fully-automated accuracy, while requiring $\sim 40\times$ less labelling effort than a purely manual reconstruction. While proof-reading is still required to correct the remaining errors in our semi-automated reconstruction, we would expect our fully-corrected semi-automated approach to remain significantly more efficient than a purely manual approach.

1.5 Publications

Existing

- A paper on our model-based method for detection of 2D neural cross-sections was accepted for oral presentation at the sixth international workshop on Microscopic Image Analysis with Applications in Biology (MIAAB: O'Reilly et al., 2011).

Planned

- A journal paper analysing our model-based methods for detection of 2D neural fibre cross-sections and 3D neurite reconstruction.

- A journal paper analysing the ultrastructure of the molecular layer of the cerebellum based on our manually reconstructed ground truth data.

- The electron microscope images and ground truth labelling for this data set will be published in the open access Cell Centered Database (CCDB) once we have published our ultrastructure paper.

Chapter 2

Reconstructing the connectome

2.1 Connectomics

In this chapter we introduce the concept of a *connectome* and discuss the challenges of reconstructing a connectome using various imaging approaches. We highlight the need for advances in automated reconstruction techniques, but defer a detailed discussion of automated reconstruction approaches to chapter 3.

2.1.1 What is connectomics?

Connectomics is the study of connectivity within the brain. As a named research field it is relatively new (Sporns, Tononi, and Kotter, 2005; Hagmann, 2005), however it is essentially the combination of anatomy and functional recording in the age of "big data". Connectomics draws together a variety of existing approaches for understanding both functional and structural neural connectivity at a range of scales.

Macro-scale

At the largest scale, Magnetic Resonance Imaging (MRI) can tell us about connectivity between different brain regions. Functional MRI (fMRI) can tell us which areas of the brain change activity levels in response to stimuli or have synchronised activity at rest. Methods have been developed to infer the *functional* connectivity between brain regions from such data. Diffusion MRI can reconstruct the bundles of myelinated axons that connect neurons in different areas, informing us about the *structural* connectivity between brain regions. Current MRI resolution is limited to voxels of 1-2 mm^3, containing $\sim$10-20,000 neurons in grey matter and potentially millions of myelinated axons in white matter. Therefore MRI can only inform us about the statistical connectivity between relatively large regions of the brain. fMRI relies on the Blood Oxygen Level Dependent (BOLD) signal related to oxygen uptake by active neurons. Therefore, it is also limited in temporal resolution to about 2 seconds. Other techniques used for investigating functional connectivity at a macro-scale include Electroencephalography (EEG),

Positron Emission Tomography (PET) and Magnetoencephalography (MEG).

In 2010 the National Institutes of Health (NIH) launched the 5 year *Human connectome project*, providing $40 million to two consortia. The first consortium (WU-Minn) plans to collect macro-scale connectomes from 1,200 healthy adults using task-based fMRI, resting state fMRI, MEG, EEG and diffusion MRI. It has recently released its first data set, containing scans for 68 individuals (Essen et al., 2013). The second consortium (MGH-UCLA) is focussing on improving diffusion MRI techniques for recovering structural connectivity. It has recently published an analysis of its initial results (McNab et al., 2013; Setsompop et al., 2013).

Micro-scale

At the smallest scale, electrophysiology and microscopic imaging can tell us about connectivity between individual neurons. Multi-electrode electrophysiology and activity-dependent fluorescence microscopy can identify small subsets of neurons that change activity in response to stimuli, informing us about the *functional* connectivity between *sparsely* sampled neurons. Electron microscopy (EM) can tell us about the *structural* connectivity between *densely* sampled neurons. In this work we focus on reconstructing the *structural* connectivity between neurons from EM images.

2.1.2 Understanding neural connectivity

In order to understand how the brain works, we need to understand how it represents and processes information. Much can be understood about the type of information processed by different brain regions by examining the connectivity between them. We can also understand the computations performed by some of these areas by mapping the functional responses of individual neurons to changes in presented stimuli. Both these techniques have been used extensively to explore how the visual system processes information, and we describe some of the insights that have been gained using these techniques below. The visual system has some key advantages for this kind of analysis. Firstly, its inputs can be easily manipulated by changing the visual stimulus presented. Secondly, it is relatively easy to physically access this area of the brain to perform functional recording. However, even with these advantages, it has only been possible to understand the detailed computations performed by the early stages of the visual system. In order to understand the computations performed by higher level processing areas of the brain, it is necessary to understand the connectivity of the neural circuits they contain. This is where micro-scale connectomics can play an important role. However, despite having being an active area of research for around 40 years, only one whole-animal connectome exists, that of the nematode worm *Caenhorabditis elegans* (*C.elegans*). However, micro-scale connectomics

has recently been combined with micro-scale functional recording of individual neurons to gain new insights into the early visual processing system. We discuss the *C.elegans* connectome and these recent analyses of local connectivity below.

Modelling visual processing

For early stages of the visual processing pathway we have been able to understand the function of cells by mapping the electrophysiological responses of individual cells directly to sensory input, creating spatio-temporal *receptive fields* (Hubel and Wiesel, 1959; Hubel and Wiesel, 1962). It has been shown that retinal ganglion cell receptive fields can be modelled by *Difference of Gaussian* or *Derivative of Gaussian* filters (Rodieck, 1965; Young, 1987), while many cells in primary visual cortex (V1) can be modelled using a Gabor-based *motion energy* model or a *Derivative of Gaussian* model (Adelson and Bergen, 1985; Jones and Palmer, 1987; Emerson, Bergen, and Adelson, 1992; Young, Lesperance, and Meyer, 2001). Additionally, simultaneous functional recording from cortical and pre-cortical neurons suggests that the receptive fields of many orientation selective V1 cells can be largely explained by the feed-forward combination of output from unoriented sub-cortical cells (Tanaka, 1983; Reid and Alonso, 1995). However, for higher-level brain areas the functional mapping between sensory input and neuronal activity becomes far too complex to measure directly. Even for the regions of visual cortex immediately upstream from V1 such mapping becomes very difficult. The mapping between the various regions involved in visual processing is reasonably well understood, and we can infer the types of information processed in each region (Felleman and Van Essen, 1991). However, in order to understand the exact computations these sub-regions perform, a detailed understanding of the connectivity between individual neurons is required.

The first and only whole-animal connectome

At the start of the 19th century an attempt was made to analyse the nervous systems of the intestinal worms *Ascaris lumbricoides* and *Ascaris megalocephala* using light microscopy (Goldschmidt, 1908; Goldschmidt, 1909). Detailed descriptions of the various neural sub-structures and bundles of neural fibres were made. However, individual neural fibres could not be resolved due to the limited resolution provided by light microscopy. It was not until after the development of the electron microscope in 1931 that individual neural fibres could be resolved and a true connectome constructed. The first whole-animal connectome was published in 1986, describing the 302 neurons and ∼8,000 chemical and electrical synapses of the nematode worm *Caenhorabditis elegans* (*C. elegans*: White et al., 1986). Although some use of computer-aided reconstruction techniques were made, the connectome was primarily reconstructed via manual annotation of ∼8,000 printed electron microscope photographs from 5 different animals. The

complete reconstruction took between 10 and 15 years to complete, with descriptions of neural sub-structures published from the same data over the previous 11 years (Ward et al., 1975; White et al., 1976; Sulston, Albertson, and Thomson, 1980).

Despite the connectome of *C. elegans* being known 12 years before its genome (published in 1998), it remains the only whole-animal connectome published to this day, although it has been further analysed and updated since (Durbin, 1987; Chen, Hall, and Chklovskii, 2006; Varshney et al., 2011; Sohn et al., 2011). Furthermore, until very recently this connectome was restricted to the hermaphrodite worm. Male worms have an additional 81 neurons, primarily located in the tail, and this final portion of the *C. elegans* connectome was not published until 2012 (Jarrell et al., 2012).

Local connectomes

Recently, several studies have reconstructed micro-scale structural connectivity of local neural circuits from electron microscope data. Some of these studies have also combined this structural connectivity with information provided by functional imaging, providing insights into the structural basis for the functional properties of neurons. Bock et al. (2011) characterised the *functional* orientation preference of 14 cells in the mouse visual cortex using calcium imaging, and characterised their structural connectivity preferences using electron microscope images. They found that the excitatory inputs to inhibitory interneurons came from cells with a broad range of orientation preferences. However, the power of their study was limited by the large number of neurites that left the imaged volume. Therefore, the dense local connectivity between the neurons of interest could not be fully recovered. Briggman, Helmstaedter, and Denk (2011) investigated the mouse retina, again combining functional calcium imaging with structural connectivity derived from electron microscope data. They focussed on the local connectivity between starburst amacrine cells and direction selective ganglion cells. The combination of the thinness of the retina and the high physical overlap between the amacrine and ganglion cells meant that a dense reconstruction of the local connectivity between 24 amacrice cells and 6 ganglion cells was possible. Briggman, Helmstaedter, and Denk demonstrated that asymmetry in the *structural* arrangement of amacrine cell inputs contributes to the *functional* direction selectivity of the ganglion cells. Takemura et al. (2013) reconstructed the structural connectivity between 379 neurons in the *Drosophila* visual system from electron microscope data. This covered an entire processing column and its connections to its neighbours. From this, they were able to identify the structural circuit underlying motion detection. Finally, Helmstaedter et al. (2013) made an extended study of the mouse retina covering 950 neurons, and examined their structural connectivity using electron microscope data. From this purely structural connectivity

they were able to make new inferences about the classification of neurons and their expected motion sensitivity. This included the discovery of a previously unknown type of bipolar cell. This was possible due to the dense nature of the reconstructed connectivity. The Takemura et al. and Helmstaedter et al. studies took advantage of automated reconstruction techniques to drastically increase the efficiency of the reconstruction process and reconstruct the connectivity between a large number of neurons. However, each reconstruction still required ∼15-20,000 person hours of manual effort to complete.

2.1.3 Methods for connectomics

Early insights into information flow in the brain were achieved from anatomical structural analyses of sparsely labelled cells (Cajal, 1894). However, in the latter half of the 20th century the capability to record the electrical activity of individual neurons (electrophysiology) appeared to make anatomy less relevant. In more recent years, *functional imaging* using fluorescence microscopy has provided a powerful complement to electrophysiology. With functional imaging, the responses of a larger number of neurons can be probed over a wider area, using calcium or voltage sensitive dyes and genetic constructs. However, very recently the idea of large scale anatomical analysis of neural circuits has regained popularity under the banner of *connectomics*. Here we briefly address the capabilities of various methods of determining micro-scale functional and structural connectivity. Later we will expand on the use of light and electron microscopy for determining structural connectivity (section 2.2).

Electrophysiology

Interest in understanding the brain as an electrical system began with the discovery by Galvani in the late 18th century that electricity could induce movement in animal muscle (Galvani, 1791; reprinted: Galvani, 1792; translated: Green, 1953). However, it was not until the recording of the first neuronal action potential ("spike") almost 150 years later that the field of electrophysiology really came into being (Hodgkin and Huxley, 1939). The key benefits of electrophysiology are unparalleled spatial and temporal resolution and extremely low noise levels, with current techniques capable of recording both spikes and sub-threshold membrane potentials in sub-cellular dendritic compartments. Key limitations of electrophysiology have been the low number of neurons that could be recorded simultaneously and the distances between recorded cells. Advances in multi-electrode recording now permit the simultaneous recording of hundreds of neurons, with the possibility of recording thousands of neurons on the horizon (Field et al., 2010; Ethier et al., 2012; Einevoll et al., 2012; Borton et al., 2013). This opens the prospect of recording a significant subset of the neurons in a cortical microcircuit. However,

the intra-electrode spacing for multi-electrode recordings is still quite large (∼30-60 μm) and therefore not all neurons in the region of the electrode are recorded. Additionally, only neurons that have a significant level of activity are recorded, leading to "silent" neurons being missed. However, the primary limitation of electrophysiological recording is that it requires a physical probe to be inserted into the neural tissue of the animal. This requirement for physical space will limit the size of probes that can be used without damaging the area being recorded from. This will in turn limit the size of the area that can be recorded from in one session. For organisms with stereotypical neural connectivity, such as *C. elegans* or *Drosophila*, recordings from many animals could be combined to make a functional electrophysiological connectome. However, neural connectivity for many organisms of interest varies significantly from subject to subject. Repeated repositioning of the probe may permit multiple recordings from the same subject. However, probe placement also causes damage to neural tissue, and therefore a whole brain electrophysiological connectome is likely to remain impossible for many organisms, including mammals.

Light microscopy

Light microscopy played an important role in the first investigations of structural connectivity by Cajal (1894). Recently, light microscopy has been developed into a powerful technique for probing the *functional* connectivity between neurons. Neurons can be stained with a variety of dyes that are sensitive to the changes in voltage or calcium levels that occur during neural activity. These stains can be targeted to particular cell types using genetic constructs, either introduced using virus injections or by developing a new strain of genetically modified animal. These techniques are therefore very powerful when applied to organisms that are genetically tractable, such as *Drosophila* or the mouse. Maisak et al. (2013) recently used such a genetically encoded calcium indicator to probe the functional properties of cells in the *Drosophila* visual system, complementing the structural analysis performed by Takemura et al. (2013). Improvements to the temporal response of such dyes and the number of neurons that can be recorded from simultaneously have only increased the usefulness of such tools. However, practical light microscopy techniques for biology are currently limited in spatial resolution by the diffraction limit. While some recent advances have improved the resolution of light microscopy beyond this limit, they are not currently suited to reconstructing the dense connectivity between neurons. The application of light microscopy to *structural* connectomics is discussed further in section 2.2.1.

Electron microscopy

Electron microscopy is currently the only technique with sufficient resolution to support the reconstruction of the dense connectivity between neurons. While there has historically been a trade-off between achieving a sufficiently high z-resolution and imaging a sufficiently large volume, recent advances have begun to address this issue. The application of electron microscopy to micro-scale *structural* connectomics is discussed further in section 2.2.2.

2.1.4 Scale of the challenge

Reconstruction effort

To reconstruct a micro-scale connectome for any organism is a massive undertaking. Table 2.1 estimates the imaging time and manual tracing effort required to reconstruct various volumes of neural tissue. Generating a volume reconstruction of a mammalian neural microcircuit is well within the resources of a single lab. All that is required is one commercially available single-beam electron microscope and a handful of tracers. Several such microcircuits have been reconstructed (Bock et al., 2011; Briggman, Helmstaedter, and Denk, 2011; Helmstaedter et al., 2013; Takemura et al., 2013). However, even stepping up to a $\sim$10,000 neuron cortical column begins to hit the limits of what can be achieved with currently available technology. A single-beam electron microscope can image a field of view of $\sim 10^6$ μm^3. Imaging a 1 mm^3 cortical column would require either 100 beams in a single microscope or for the tissue to be split into 100 sub-volumes for imaging without losing any tissue at the sub-volume boundaries. The former is possible with the latest research microscopes in development. A 196-beam microscope has been demonstrated, but for etching rather than imaging (Mohammadi-Gheidari, Hagen, and Kruit, 2010), while Zeiss are developing a 61-beam microscope with the capability to image all 61 beams (Perkel, 2013). Lossless splitting of neural tissue has been demonstrated using *hot knife microtomy* by Hayworth et al. (2010), though only as a proof of concept. Even if the imaging challenge is solved, a cortical column reconstruction is likely to be limited to a skeleton tracing as a volume reconstruction would require thousands of person years of manual tracing.

To reconstruct a human connectome is well beyond the reach of current global imaging and tracing resources. Even with 1,000 196-beam microscopes, imaging alone would take 350 years and a skeleton tracing would require 200 million person years of effort. To image a human brain in 10 years would require significant advances in multi-beam imaging. However, even with such an extreme advance in imaging capability, automated or semi-automated tracing techniques would need to be developed to support a similarly extreme increase in skeleton trac-

ing speed. Even with 10,000 tracers, a 2,000-fold increase in tracing speed would be required. Even reconstructing a mouse connectome within 10 years would stretch the resources of a large consortium today. Imaging would require 250 single-beam microscopes and a skeleton tracing would require almost 7,000 tracers.

	Microcircuit	Cortical column	Mouse brain	Human brain
Volume	10^6 μm^3	1 mm^3	500 mm^3	1.4×10^6 mm^3
Imaging (multi)	2.2 hours	13 weeks	130 years	350,000 years
Imaging (single)	18 days	49 years	25,000 years	69 million years
Skeleton tracing	7.1 weeks	140 years	69,000 years	190 million years
Volume tracing	6.9 years	6,900 years	3.4 million years	9.6 billion years

Table 2.1: Estimates of imaging time and manual tracing effort required to reconstruct **(i)** a cortical microcircuit, **(ii)** a cortical column, **(iii)** a mouse brain and **(iv)** a human brain. Reconstructing a microcircuit or cortical column is within the resources of a single lab. However, even a skeleton connectome for a mouse brain would stretch the resources of a large consortium. Reconstructing a human connectome of any kind is well beyond the reach of current global imaging and tracing resources. Total reconstruction effort depends on the number of imaging beams and whether a skeleton or volume tracing is required. Times are for a single scanning block-face electron microscope and a single human tracer. Tracing is trivial to parallelise with additional tracers. Lossless splitting of a tissue sample for imaging on multiple microscopes is less easy, but should be possible using *hot knife microtomy*. Imaging estimates use a voxel scanning rate of 100MHz per beam with an isotropic voxel resolution of 18.6 nm (from our FIBSEM), and the multi-beam estimate assumes 196 beams (Mohammadi-Gheidari, Hagen, and Kruit, 2010). Tracing estimates use rates from Helmstaedter, Briggman, and Denk (2011).

Comparison with the human genome project

It is interesting to compare progress in the field of connectomics with that in the field of genomics. This allows an appreciation of the scale of the problem. The first genome was published in 1977, describing the ∼5,400 nucleotides comprising the DNA of the bacteriophage $\phi X174$ (Sanger et al., 1977). By 1985, the genomes of multiple bacteria with ∼40-50,000 nucleotides had been sequenced, and early work was underway to sequence the genome of *C. elegans*, which would be published over 10 years later (Consortium, 1998). This prompted the now famous *Santa Cruz workshop* in May 1985 (Sinsheimer, 1989), which led to the launch of the $3 billion *human genome project* in 1990 (DoE, 1991). This project involved two major programs. The first was a public project, with 600 sequencing machines distributed in labs across the world. This was later joined by a private project based at Celera, with 300 auto-

mated sequencing machines. After 10 years of concerted effort, the initial human genome was published in 2001 (Consortium et al., 2001; Venter et al., 2001). During the 10 years of the human genome project, sequencing speeds had improved by less than 200-fold. However, over the following 10 years both sequencing speeds and costs had improved over 100,000-fold. In 2012, the first comparative analysis of over 1,000 human genomes was published (Consortium, 2012). This represents an astonishing advance in sequencing capability, and it is unlikely that similar progress would have been made in the absence of the human genome project. If current connectomics imaging and tracing speeds were to experience the same ∼20 million fold improvement as genome sequencing speeds have over the past 20 years, reconstructing the human connectome would become achievable. Imaging time would fall to ∼3 beam years, skeleton tracing time to ∼10 person years and even volume reconstruction would be within reach at ∼480 person years. There is an argument to be made that reconstructing the human connectome is a significantly more difficult challenge than reconstructing the human genome. The human genome contains ∼3 billion base-pairs organised in 20-30,000 genes with relatively simple 1D geometry. In contrast, the human connectome has ∼86 billion neurons organised into millions of highly interconnected microcircuits, arranged in a complex 3D geometry. However, even with the benefit of a similar 20 million fold improvement in reconstruction speed, reconstructing a skeleton human connectome would still be over 3,000 times slower than reconstructing a human genome. Therefore, it could equally be argued that this difference already reflects the additional difficulty.

If improvements in electron microscope imaging techniques improve at their current pace, imaging a human brain might become feasible for a large consortium. However, the lack of reliable automated or semi-automated tracing techniques is a clear bottleneck in the reconstruction process. Such techniques are under increasingly active development (section 3) and, with the recent announcements of billion-dollar brain projects by the E.U. (The Human Brain Project) and the U.S. (BRAIN), the research effort into reconstruction techniques for connectomics is likely to increase.

2.1.5 Focus on cerebellar parallel fibre reconstruction

The human brain contains ∼86 billion neurons, and ∼80% of these are located in the *cerebellum* (Azevedo et al., 2009). The vast majority of the ∼69 billion neurons in the cerebellum are *granule cells*, which form the input layer of the cerebellum and receive a range of motor and sensory inputs. The granule cell axons ascend through the cerebellum to the *molecular layer* before making a single "T" branch and running parallel to the cerebellar surface for several millimetres, making no further branches. These *parallel fibres* form the primary input to the large,

highly branching *Purkinje cells* that provide the sole output from the cerebellum. Accurately reconstructing these *parallel fibres* is therefore crucial in order to understand the circuitry of the cerebellum. Their vast number, long length and small diameter make this a challenging problem. However, their parallel nature means that imaging can be performed such that most fibres run approximately perpendicular to the image plane. As a result their cross-sections in each image will tend to be reasonably convex, and cross-sections of a single fibre in successive images will tend to overlap significantly. Both these properties lend themselves to a 2+1D approach, where fibre cross-sections are identified in each image independently and then combined into 3D fibres in a separate step. Several studies have taken such a 2+1D approach to the reconstruction of neural fibres, and this is the approach we take in this work. Figure 2.1 provides an overview of our long-term goal. By reconstructing cerebellar neurons from electron microscope images, we hope to be able to reconstruct and analyse the information processing circuits in the cerebellum. However, for this work, we focus on the restricted problem of reconstructing cerebellar *parallel fibres* only.

2.2 Imaging for connectomics

2.2.1 Light microscopy

Breaking the diffraction limit

Conventional fluorescence microscopy techniques such as confocal and two-photon microscopy are diffraction-limited to a resolution of approximately 200 nm in the image plane and 450 nm axially. At their smallest, neural fibres can be as thin as 90 nm in diameter, with a membrane thickness of 10-20 nm in grey matter. This means that diffraction-limited techniques do not have sufficient resolution to distinguish adjacent fibres. However, some recently developed fluorescence imaging techniques have managed to achieve resolutions well below the diffraction limit, approaching that required to resolve the thinnest fibres. Structured illumination microscopy (SIM; Gustafsson et al., 2008) uses interfering patterns of illumination to achieve a resolution of approximately 100 nm in the image plane and 280 nm axially. Stochastic optical reconstruction microscopy (STORM; Rust et al., 2006) and photo-activated localisation microscopy (PALM; Shroff et al., 2008) rely on exciting only a small subset of fluorophores on each scan. If a small enough subset is activated in each frame, almost all fluorophores will be separated by a large enough distance to resolve as individual points. Fitting the known point spread function of the fluorophore to each point can result in resolutions of approximately 20 nm in the image plane and 50 nm axially. Stimulated emission depletion microscopy (STED; Wildanger et al., 2009) uses a “doughnut” shaped beam to quench the edge of a diffraction-limited spot prior to

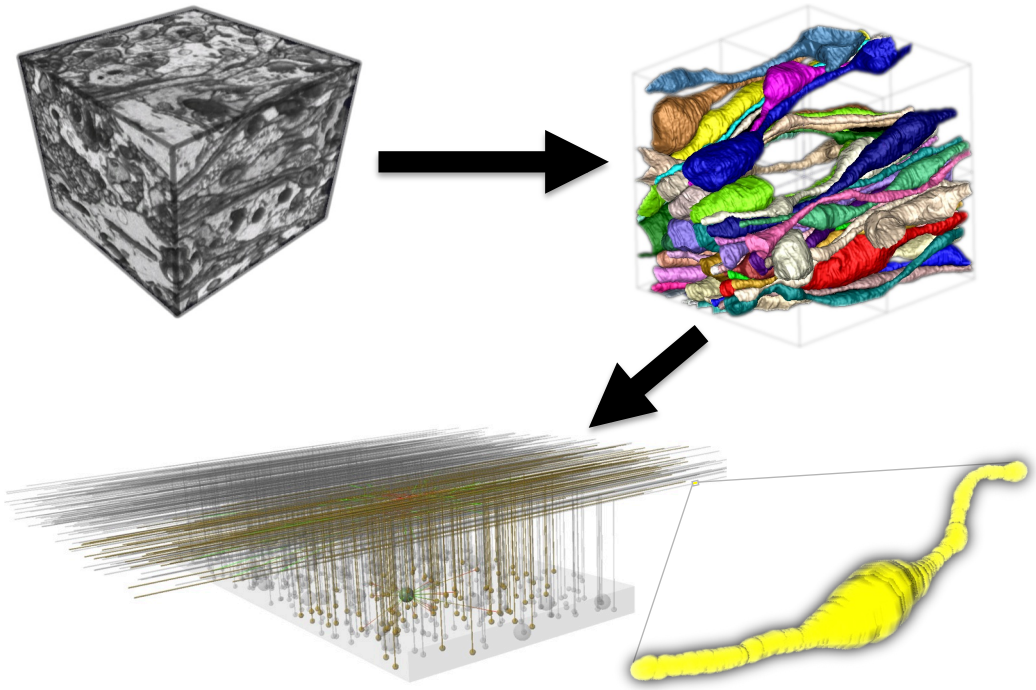


Figure 2.1: From electron microscope images to cerebellar model. The eventual goal is to take a 3D electron microscope image of the cerebellum (top left), reconstruct all the neurons in the region (top right) and generate a computational model of the cerebellum (bottom left). However, for this work we focus on reconstructing only the parallel fibres in the molecular layer of the cerebellum (bottom right). Reconstructed neurons (top right) generated using *itkSnap* (Yushkevich et al., 2006). Image of cerebellar granule cell layer network model (bottom left) generated using *neuroConstruct* (Gleeson, Steuber, and Silver, 2007) and adapted with permission from neuroConstruct website. *Source: http://www.neuroconstruct.org/models/images/GranCellLayer/large.png*. This image is licenced under the same CC-BY terms as the thesis.

fluorescence, shrinking it below the diffraction limit and achieving resolutions of up to 20 nm in the image plane and 30 nm axially. However, none of these techniques are yet ready for routine use. Additionally, limitations on achievable fluorophore density and susceptibility to photo bleaching mean that reconstructing a densely labelled volume using single-colour fluorescence imaging currently remains impossible.

Multi-colour fluorescence microscopy

Recent advances in large-scale multi-colour labelling such as *Brainbow* (Livet et al., 2007) provide a potential means of overcoming some of the issues with optical resolution and fluorophore density. In the Brainbow scheme, each neuron expresses a random level of each of four spectrally separable fluorophores. This provides additional information to resolve ambiguities due to the constraint that any reconstruction must maintain a consistent colour within each neuron. However, such techniques are still in their infancy and have issues with palette size, consistency of label intensity and photo-bleaching that mean they are not yet ready for large scale

reconstruction efforts. The dream would be to combine improved Brainbow labelling with improved sub-diffraction limit microscopy. With this combination, fibres might be reconstructed along most of their length simply by grouping pixels by colour, with manual intervention only required when two fibres of the same colour overlap. Synapses in different neurons could also be marked with different colours and functional synapses between two neurons identified from the colour combinations observed at fibre intersections. It has even been proposed that, given sufficient Brainbow colours and resolution, only the synapses would need to be marked to recover full neural connectivity (Mishchenko, 2010). However, this dream relies critically on the achievable number of spectrally separable Brainbow colours. It is unlikely that this will be sufficient to use a synapse-only labelling approach for the reconstruction of a cortical column. In fact, given the likely limitations of the Brainbow palette, it is probable that even the combination of sub-diffraction limit microscopy and Brainbow will still result in many locations where some ambiguity about fibre identity remains.

Regardless of the eventual limits of fluorescence microscopy and Brainbow-like techniques, they are not currently suitable for large scale dense reconstructions of neural circuitry. Therefore, for the immediate future at least, electron microscopy remains the only method capable of resolving the smallest neural fibres. Even if sufficient Brainbow colours and fluorescence microscopy resolution were achieved, electron microscopy will likely still be the best option for reconstructing neural circuits in the many animals (including humans) where the genetic manipulations required for Brainbow are not possible. However, a promising future avenue for genetically tractable animals would be to image the same volume using sub-diffraction limit Brainbow fluorescence imaging *and* electron microscopy. Such a multi-modal approach would provide much richer information to automated reconstruction techniques. Potentially Brainbow inconsistencies could be used to automate the identification of locations where human intervention was required and electron microscopy would provide sufficient detail for the unambiguous manual reconstruction of these (hopefully sparse) problem locations.

2.2.2 Electron microscopy

Tissue staining

All electron microscopy relies on selectively staining some features in a biological sample with electron dense material. This provides contrast between the stained and unstained features. In all cases, cellular membranes are preferentially stained. *Classical* staining marks both external and internal membranes (e.g. mitochondria, ribosome, filaments etc.). An alternative *extracellular* stain marks only cell outer membranes and extracellular space. While the extracellular stain makes the tracing of neural fibres easier, it makes the identification of synapses unreliable,

except in special cases where the geometry of inter-neuron contacts can be reliably used for synapse identification (e.g. retinal synapses; Briggman, Helmstaedter, and Denk, 2011). The classical stain provides sufficient detail for human experts to unambiguously trace neural fibres and identify synapses, but the staining of additional intracellular "clutter" makes automated reconstruction more challenging.

Electron microscope images can be taken in two imaging modes, Transmission Electron Microscopy (TEM) and Scanning Electron Microscopy (SEM).

Transmission electron microscopy (TEM)

TEM consists of taking thin slices of tissue, shining a beam of electrons through it, and collecting the electrons that are transmitted through the material. Electron dense areas scatter electrons strongly. Therefore, many fewer electrons are transmitted through these parts of the material, and they appear dark in the image. The resolution of TEM in the imaging plane can be as high as 2 nm. However, the z-resolution is generally limited to about 50 nm due to the necessity of cutting slices that are thick enough to remain intact during handling. This relatively poor z-resolution means that the cross-sections of fibres not running perpendicular to the imaging plane move significantly between slices, making them difficult to follow through the volume. The latest automated collection techniques now permit slices to be as thin as $\sim$25 nm (Schalek et al., 2012), although this is still borderline for tracing the thinnest fibres without error. Additionally, fibre membranes become increasing blurred as a fibre's direction of travel deviates from perpendicular. This is due to the fact that the transmitted electrons effectively average the position of the membrane throughout the thickness of the slice. This blurring makes individual fibre cross-sections harder to identify, and thin fibres running in or close to the imaging plane can even be missed entirely. However, recent work on combining tomography with TEM has shown promise in improving the effective resolution of TEM (Veeraraghavan et al., 2010). More importantly, even with the "thick" slices used for TEM, damaged slices are common and the resulting missing cross-sections make the 3D reconstruction problem even harder. Many slices that are not completely lost acquire tears and folds when cut, making the process of aligning successive slices to form a 3D volume challenging. It is telling that there is a large body of research dedicated to the alignment of TEM images into 3D volumes.

Scanning electron microscopy (SEM)

In SEM, the tissue is imaged by detecting electrons reflected from the surface of the sample. Most of these are *secondary* electrons, which report the surface topology. However, a small proportion of the electrons are scattered directly backwards from the electron dense regions. With a suitable detector, these *back-scattered* electrons can be isolated from the more numerous

secondary electrons. When this is done, electron dense regions such as membranes appear bright on the raw imagery. However, to maintain a common look for EM imagery, SEM images are inverted by convention so that electron dense material appears dark as for TEM. These back-scattered electrons represent a much smaller proportion of the imaging beam than the transmitted electrons used in TEM. As a result, more electrons are required for the same level of contrast, which can be provided by imaging with a more intense beam or imaging each pixel for a longer period of time. However, one of the advantages of SEM over TEM is that the penetration of back-scattered electrons can be limited by reducing their energy. As a result, SEM is effectively a surface imaging technique, and therefore not subject to the blurring effect of TEM caused by averaging electron density through the slice. However, the key advantage of SEM is that is can be used to image the surface of a block of tissue. This is discussed further below.

In addition to the two imaging modes, electron microscope images can also be taken using two imaging processes, *serial section* imaging and *serial block face* imaging.

Serial section imaging

TEM images can only be taken by slicing tissue sections from a sample prior to imaging. In order to reconstruct a large volume of tissue, serial sections need to be taken from the same tissue block. Conventionally, multiple manually cut 50 nm thick slices are imaged in turn. However, the Automatic Tape-collecting Ultramicrotome (ATUM) in development at Harvard (Schalek et al., 2012), automates the slicing and collection of sections. Slices are collected on a supportive tape and can therefore made thinner ($\sim$25 nm) without increasing the risk of damage. Making sections prior to imaging permits both fluorescence and EM images to be taken of the same volume, as well as parallel imaging using multiple electron microscopes. This makes ATUM a good candidate for applications where large field of view or high throughput is required. While ATUM permits imaging with both TEM and SEM, the achievable z-resolution is limited by the slice thickness and this is borderline for tracing the thinnest fibres without error.

Serial block face imaging (SBFSEM)

As SEM is a surface imaging technique, it is not necessary to slice the sample prior to imaging. The surface of the sample block can be imaged and then a thin slice removed prior to taking the next image. Because these slices are discarded, they can be made thinner than with serial section imaging. Therefore, z-resolution is limited only by the minimum slice thickness that can be reliably removed from the face of the sample. Additionally, the risk of sample damage associated with sectioning prior to imaging is removed. Currently, these slices are removed

either with a diamond knife (Denk and Horstmann, 2004) or a focussed ion beam (FIBSEM; Knott et al., 2008). These approaches can remove slices as thin as 25 nm and 10 nm respectively. Its superior z-resolution makes FIBSEM the only technique capable of tracking the smallest fibres running in arbitrary directions relative to the imaging plane. However the maximum area for ion beam milling at high resolution is limited to an area of approximately 100×100 μm, limiting the size of the volume that can currently be imaged with this technique.

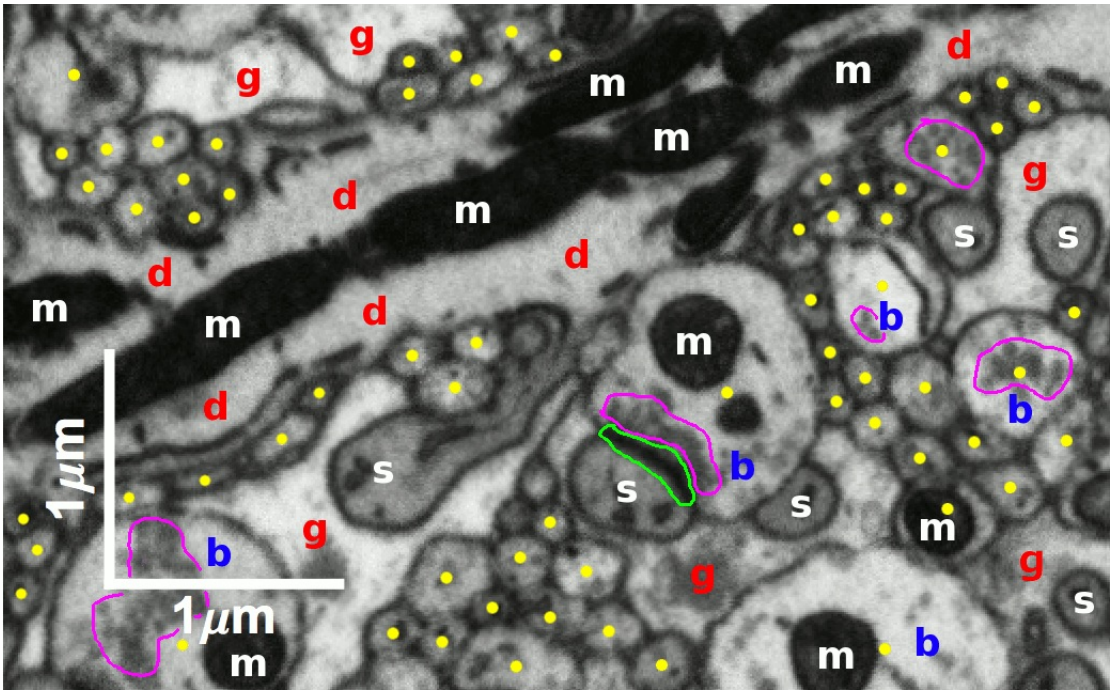


Figure 2.2: Example electron microscope image of the molecular layer of the mouse cerebellum, highlighting key neural structures. Yellow dots mark the approximate centres of axon cross-sections. Most of these will be parallel fibres (granule cell axons) and will run perpendicular to the image plane. However, some will be interneuron axons and will run at a wider range of angles. There is a single Pukinje cell dendrite cross-section in this image and its extent is marked with the red letter *d*. This dendrite contains lots of mitochondria (white *m*), which convert stored energy into a form that is useful for the cell. Purkinje cell spines are marked with a white letter *s*. These are thin protrusions from a Purkinje cell dendrite that make synapses with parallel fibres. Most of the spines in this image belong to the Purkinje cell dendrite seen in the image. Periodically, parallel fibres will swell to form a pre-synaptic bouton, and these are marked with a blue letter *b*. Boutons often contain mitochondria (white *m*) and vesicles (extent marked by purple boundaries). Vesicles are small spherical capsules formed of cellular membrane that contain neurotransmitters. When a synapse is activated, these dock with the cell membrane at the synapse and release their contents, activating receptors in the post-synaptic cell. Most synapses made by parallel fibres are with Purkinje cell spines, and one can be seen near the centre of this image. The characteristic post-synaptic density that indicates a synapse is highlighted in green. The space between neurons is filled by support cells called glia (marked by a red letter *g*). This sample has been stained using the classical *intracellular* stain. Segmentation of cellular cross-sections is very difficult in such samples, as many of the intracellular structures (e.g. vesicles and mitochondria edges) look very similar to extracellular membrane when only a small local neighbourhood is considered.

Chapter 3

Image segmentation for connectomics

In this chapter we review the range of approaches that have been applied to the problem of reconstructing neurites from electron microscope (EM) images. These approaches can be broadly grouped into bottom-up *pixel-based* methods and top-down *contour-based* methods. We discuss some of the issues with both these approaches, and suggest how our circle-based approach might address some of these issues. We also discuss semi-automated reconstruction approaches. Finally, we review measures used to quantify reconstruction accuracy, before describing the measures we have chosen to use for this work.

3.1 Pixel based approaches

The majority of EM reconstruction approaches explicitly attempt to distinguish between *membrane* pixels that form the boundaries between cells and *non-membrane* pixels that form the interior of cells. In EM images, the membrane pixels that comprise the boundaries of cells are significantly darker than most pixels representing intracellular space. However, when using a *classical* stain, intracellular structures are also stained as darkly as cell membranes. Some studies that focus on region merging or linking cross-sections across slices simply threshold the grey-level EM image to separate light and dark pixels, perhaps applying some filtering to enhance the boundaries before or after thresholding (Jurrus et al., 2008; Yang and Choe, 2009). However, in most cases such a simple approach is likely to be insufficient to produce a good quality segmentation. Therefore most studies use machine learning to classify pixels as either *membrane* or *non-membrane*. We discuss some of the most commonly used machine learning techniques used to classify pixels below. We then discuss the types of features commonly used as the input to such classifiers, as well as the issue of distinguishing between cell membrane and intracellular structures when using a *classical* stain. Finally, we discuss techniques used to group pixels into either 2D neurite cross-sections or 3D neurite segments, and the techniques used by 2+1D approaches to link neurite cross-sections across slices.

3.1.1 Techniques for learning to predict pixels

Artificial neural networks (ANNs)

The key component of artificial neural networks is the *perceptron* (Rosenblatt, 1958). It outputs the result of passing a linear weighted sum of its inputs through an *activation function*. Training a perceptron requires the use of a suitable *learning rule* to set its weights from a collection of training inputs and target outputs. With binary target outputs and a strongly non-linear activation function, a perceptron can learn to perform *binary classification*. With real-valued target outputs and a suitable selection of activation function, a perceptron can learn to perform *linear* or *logistic* regression. Multiple perceptrons can be coupled together in a network to form a *multi layer perceptron* (MLP), which is capable of learning a large range of non-linear functions. Although these MLPs were known about for some years prior, it was not until the development of the *backpropagation* algorithm that they were able to be effectively trained. This algorithm was most famously introduced by Rumelhart, Hinton, and Williams (1986), although it had been previously reported by Werbos (1974) and LeCun (1985).

Several studies from the Scientific Computing Institute (SCI) at the University of Utah have trained multiple MLPs connected in series to identify extracellular membrane (Jurrus et al., 2009b, 2010, 2013; Seyedhosseini et al., 2011). In such serial networks, the first stage only receives input from the EM image, but subsequent stages combine the image input with the output of the previous stage. By passing the output of each stage to the next in the series, each stage combines information from a wider spatial context. The networks reported in the different studies differ primarily in the type of image features used and the methods used to segment the membrane probability map output by the network.

Convolutional neural networks (CNNs)

Convolutional neural networks are a reformulation of MLP neural networks. One property is that they take the raw image as input, rather than the output of a pre-determined set of image filters. In this way, they learn the most appropriate image features for the task. However, learning initial image features this way can equivalently be done by using an MLP NN that takes image patches as input. This is achieved by applying the patch-based MLP NN to overlapping image patches (i.e. convolving it with the image). The key difference between a CNN and a patch-based MLP NN is that the connections between the layers of the CNN are also convolutional in nature, taking overlapping "image" patches of the previous layer. In contrast, the connections between layers in an MLP NN are restricted to single scalar weights. Strictly speaking, any CNN can be equivalently constructed as a more complex MLP NN with the convolutional nature enforced by constraints requiring subsets of connection weights to be shared. However,

as additional hidden nodes are added to a CNN, the representation of the equivalent MLP NN becomes more complex and less intuitive to interpret.

CNNs were formally introduced by LeCun et al. (1989), when he demonstrated how to apply the *backpropagation* algorithm to such networks. However, the concept of neural networks with a convolutional structure had been around for some time (Fukushima, 1980), inspired by the receptive field structure suggested for the mammalian visual system by Hubel and Wiesel (1962). Until recently MLPs with many hidden units and/or many layers could not be effectively trained in a reasonable amount of time. However, with the advent of Graphical Processing Units (GPUs) capable of very fast parallel processing, effective training of large, deep networks has become possible (Cireşan et al., 2010).

CNNs have been applied to the problem of segmenting neurons in EM images by groups at the Massachusetts Institute of Technology (MIT: Jain et al., 2007, 2010; Turaga et al., 2009, 2010) and the Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (IDSIA: Cireşan et al., 2012; where CNNs are referred to as Deep Neural Networks). The networks from both groups have performed very well, achieving state of the art performance on tissue with *extracellular* staining when trained to optimise pixel classification accuracy. The MIT group have also extended the approach to train CNNs to directly optimise the pixel pair based *Rand index* (Turaga et al., 2009) and the topological *warping error* (Jain et al., 2010), with the most recent work reporting excellent results on tissue with *classical* intracellular staining. Both measures are more representative of object-level segmentation performance than the pixel classification accuracy. However, training the network on these more representative measures is significantly slower than training it on the pixel accuracy measure. For a given amount of time spent training a network, there is therefore a trade-off between the complexity of the measure optimised and the size of the network and volume of training data that can be used.

Random forests (RFs)

Random forest classifiers (Breiman, 2001) are a popular approach to building a strong classifier from many weak classifiers. Random forests are constructed by combining many *binary decision trees*. Each node of each tree splits the data using a learned threshold on a single input feature, selecting the feature and threshold that best separates a subset of the training data. The training data for each tree is limited to a random subset of the full training data. The features considered for splitting the data at each node are also limited to a random subset of the available input features. This randomness ensures that correlations between the trees of the forest are kept low. As a result, combining the output of the trees results in much better classification performance than optimising a single tree. Section 6.4.3 describes random forests in more

detail.

Much of the work on using random forests to segment neurons in EM images has been done at the Heidelberg Collaboratory for Image Processing (HCI). This group have trained a random forest pixel classifier to use a wide range of image features and have achieved state of the art membrane classification performance on tissue with *extracellular* staining (Andres et al., 2008, 2012). Their random forest classifier and many of their image features features are available via a C++/Python image processing library (VIGRA: Köthe, 2012) and an interactive segmentation program (*ilastik*: Sommer et al., 2011). In this work, we benchmark the performance of our circle-based algorithm against that achieved by *ilastik* on our image data (section 6.7). Other groups have also used random forests pixel classifiers to classify extracellular membrane (Kaynig, Fuchs, and Buhmann, 2010b; Laptev et al., 2012) and intracellular mitochondria (Giuly, Martone, and Ellisman, 2012).

Other machine learning techniques

Several other machine learning methods have been used to learn pixel classifiers for identifying neurons in EM images. Examples include *k-means clustering* (Lucchi et al., 2010), *support vector machines* (SVMs: Glasner et al., 2011; Lucchi et al., 2010) and *boosted decision stumps* (Venkataraju et al., 2009).

3.1.2 Image features

A wide range of image features are used in segmentation algorithms. These encode the structure of the image in the vicinity of a pixel in various ways. We discuss some of the most common feature types below.

Patch-based features

These features consider the local neighbourhood of a pixel, and can encode a variety of local image properties. Many algorithms make use of features based on the first and second order derivatives of the image. The Basic Image Features (BIFs) we use for our reconstruction algorithm fall within this category (section 5.4.1). Other popular patch-based features include Gabor filters, the structure tensor and the Sobel edge detection filter. Several studies also make use of features based on the statistics of image patches, such as the mean and variance of the image luminance. A recent trend is to use a wide variety of patch-based filters in a single algorithm (Andres et al., 2008, 2012; Laptev et al., 2012). While many patch-based image features perform mathematically defined operations on the local image neighbourhood, it is also possible to define arbitrary features. Knowles-Barley et al. (2011) use filter banks containing explicit examples of membrane image patches, while the *convolutional neural networks* discussed in

3.1.1 learn a set of arbitrary image patches that are most useful for predicting the presence of membrane.

Line-based features

These features consider the image properties along a line passing through a pixel. Examples include *ray features* (Smith, Carleton, and Lepetit, 2009) and *radon-like features* (Kumar, Vazquez-Reina, and Pfister, 2010). *Ray features* consider a set of line segments passing through a pixel and terminated at the nearest boundary pixel. They then consider the properties of these end points, such as the distance of the end points from the pixel and the image gradient at the end points (Lucchi et al., 2010). *Radon-like* features extend *ray features* to add sampling or aggregation of image properties along the length of each segment, rather than only at the end points. Although ray and radon-like features are line-based rather than patch-based, they both make use of patch-based features. Ray features explicitly compute patch-based gradient features at each line segment end point, while both approaches use patch-based features to generate the boundary maps they rely on.

Stencil-based features

These are an extension of patch-based features to encompass a wider image context. Stencil-based approaches sample pixels from a wider image neighbourhood, reducing the density of sampling as the distance from the centre of the stencil grows. This non-uniform sampling permits the consideration of information from pixels in a larger neighbourhood when compared to a patch-based feature using the same number of pixels. The main studies making use of stencil-based features for connectomics are from the Scientific Computing Institute (SCI) at the University of Utah (Jurrus et al., 2009b, 2010, 2013; Seyedhosseini et al., 2011).

3.1.3 Intracellular structure identification

One of the key difficulties when identifying neurons in *classically* stained electron microscope images is distinguishing between the external membrane forming the boundaries between cells and intracellular structures. When an *extracellular* stain is used this problem does not exist, as only the external membrane is stained. However, for most brain regions, the unambiguous identification of synapses then becomes impossible. A *classical* stain is necessary to ensure that synapses can be unambiguously identified. However, intracellular structures such as mitochondria and synaptic vesicles are then stained alongside the external membrane. Given the relatively local context considered by many image features, successfully distinguishing between external membrane and intracellular structures can be difficult. In our model-based algorithm, we attempt to address this issue by only considering features in the neighbourhood of the ex-

pected location of the external membrane. However, this approach is not foolproof, and being able to reliably exclude intracellular structures would be expected to improve the performance of our algorithm. Several studies have proposed methods to identify synapses (Knowles-Barley et al., 2011; Kreshuk et al., 2011) and mitochondria (Giuly, Martone, and Ellisman, 2012; Knowles-Barley et al., 2011; Seyedhosseini, Ellisman, and Tasdizen, 2013).

3.1.4 Pixel grouping

The output of most pixel classifiers is a *membrane probability map* that predicts how likely each image pixel is to be part of an external membrane separating two neurons. Such maps require further processing to convert them into reconstructed neurite segments. The simplest approach is to threshold the probability map to generate a binary labelling. The *connected components* algorithm can then be used to cluster non-membrane pixels into isolated groups separated by membrane pixels. A more sophisticated approach is to use the *watershed* method, which treats the membrane probabilities as heights in a landscape, and clusters non-membrane pixels into regions that would be in the same "lake" given a certain water level (Beucher and Lantuéjoul, 1979). A third approach is to use a graph cut approach (e.g. Vu and Manjunath, 2008). In addition to considering the probability for each pixel, graph cut methods can include additional terms to enforce prior knowledge about the expected segmentation (e.g. smoothness of labelling). Some methods use the results of a watershed or graph cut clustering as their final segmentation. However, other methods add an additional region merging step. By selecting appropriate parameters for the watershed or graph cut stage, an *over-segmentation* can be achieved. In an *over-segmentation* most true neurite segments will be split into multiple pixel clusters in the algorithm output. This will produce a poor quality segmentation, but the likelihood of a pixel cluster merging two true neurite segments will be low. A second stage classifier can then be trained to merge the over-segmented pixel clusters into accurate neurite segments. Several studies have used this approach, learning an optimal merging strategy based on the features of clusters and the junctions between them (Andres et al., 2008, 2012; Lucchi et al., 2010). Liu et al. (2012) take this approach a step further and consider all possible watershed segmentations when considering which regions to merge.

3.1.5 Linking cross-sections across slices

Many automated reconstruction algorithms take a 2+1D approach to segmentation. In this approach, neurite cross-sections are first found independently in each slice. These 2D cross-sections are then linked together across slices to form 3D neurites. In many studies, the method used to link cross-sections across slices is to treat the set of cross-sections as nodes on a graph.

Nodes in adjacent slices are connected by edges, and the weights assigned to these edges are dependent on the consistency between pairs of cross-sections. Cross-sections are joined across slices either by finding a set of minimum cost paths through the graph (Jurrus et al., 2008, 2013), or by performing a hierarchical clustering (Kaynig, Fuchs, and Buhmann, 2010a). Vazquez-Reina et al. (2011) extend this approach further, by considering a range of possible 2D segmentations in each slice when evaluating the optimal 3D linkage between cross-sections, in a manner similar to Liu et al. (2012).

3.2 Contour-based approaches

Almost all model-based approaches applied to EM neural reconstruction use a version of the level-set active contour technique. This models each surface as a flexible contour, with the topology of an image-based energy function causing the contour to move and deform such that it settles on the boundary of an object in the image. The algorithm permits the splitting and merging of contours, so branching in neural fibres should be handled elegantly, which is a strong advantage of this approach. However, the performance of this technique depends strongly on the chosen energy function and the initial location of the contours. The simplest approach uses an energy function purely based on the image. For example, to fit an isolated neuron in a fluorescence image, the energy function might have a minimum where the intensity gradient was maximal. In this case a contour could be seeded anywhere inside the neuron and it would eventually grow to trace its boundary. In practice the algorithm might have trouble bridging small gaps of low fluorescence without also extending beyond the boundaries of the neuron. This can be partially countered by making the propagation speed of each point on the contour proportional to image intensity. Within the neuron, the contour would rapidly spread, and small gaps would eventually be crossed. "Leakage" outside the neuron would be limited to a distance of the order of the largest gap crossed.

However, contours with energy terms based purely on the image perform very poorly with *classically* stained EM data, due to the dense packing of fibres and the large amount of intracellular clutter. It is almost impossible to tweak the propagation rate of the contour to ensure that intracellular clutter is passed over, without also permitting the contour to easily cross cell membranes. A refinement of the technique takes advantage of prior knowledge about the shape of cell membranes in EM data. Fibre cross-sections can be expected to have membranes that vary smoothly, with no sharp changes in direction. This knowledge of the membrane geometry can be built into the energy function through the incorporation of an additional *contour stiffness* term that limits both the absolute curvature of the contour and its rate of deformation. This

refinement drastically improves the performance of this technique on EM data (Jurrus et al., 2009a; Macke et al., 2008). A further refinement to the active contour model replaces each contour with a pair of interacting contours. Another addition is made to the energy term to add a force between each pair of contours that attracts them at a long scale but repels them at a very short scale. With this additional force, the contour pair finds the membrane more accurately so long as one contour is initialised within the membrane and the other initialised to surround the membrane (Vazquez-Reina, Miller, and Pfister, 2009). Note that this improvement in performance seems to be very reliant on the quality of the initial contour positioning. This is the fundamental issue with all contour-based approaches.

3.3 Issues with pixel and contour based approaches

We have discussed two main approaches for reconstructing neurites from EM images. *Pixel-based* methods use *local* image features and machine learning to predict whether pixels are *membrane* or *non-membrane*. These features are often fast and simple to compute but can only consider limited spatial context. While several studies have explored methods to increase this context, none have fully solved the problem. As a result, locally weak membrane evidence at a cell boundary will often result in a gap in the detected boundary, erroneously joining two cells. Conversely, locally strong membrane evidence from intracellular structures will often result in erroneous detection of boundary pixels within cell interiors. These errors in boundary pixel detection will often result in the subsequent segmentation containing structures with highly non-convex boundaries, even though the true boundaries for the neural cross-sections we are interested in are mostly convex. In contrast, *contour-based* approaches can take into consideration long-range regularities in boundary structure by attempting to find the set of boundary contours that are well supported by the image data, while exhibiting the regularities observed in the ground truth data. Such regularities can include both geometrical constraints (e.g. convexity) and interaction constraints (e.g. limiting overlap between contours). However, the optimal solution cannot be directly constructed and must instead be searched for. The solution space of possible sets of contours is too vast to exhaustively evaluate and can only be searched by local refinement. It is likely to contain many local optima and so the quality of the found "optimal" solution is highly dependent on where the search starts. Thus, while this method is effective for propagating a known good set of contours to an adjacent slice, it is much less effective when a good quality initialisation is unavailable.

We model the cross-sections of neurites as circles, which addresses the key issues with existing pixel-based and contour-based approaches. We evaluate the image evidence for each

circle within an annular region around its perimeter. This results in the consideration of evidence from a larger context than most pixel-based methods, and permits us to integrate image evidence over the entire boundary of a fibre cross-section in a similar manner to contour-based methods. The use of circles as our model of fibre cross-sections results in a drastic reduction in the number of degrees of freedom compared to contour-based methods. This permits us to evaluate the evidence provided by the image for a full range of candidate circles at each pixel. This exhaustive evaluation of the solution space avoids the problem of local minima associated with contour-based methods. Our approach is discussed further in chapter 5.

3.4 Semi-automated approaches

While research into automated reconstruction algorithms has been extensive, no fully-automated method currently produces acceptably accurate reconstructions without substantial human proof-reading and correction. There is therefore great interest in developing semi-automated approaches that make the most efficient use of manual reconstruction effort. Several interactive semi-automated reconstruction programs have been developed, and we summarise some of them here. Sommer et al. (2011) incorporate many aspects of the approach described by Andres et al. (2008) into an interactive segmentation program called *ilastik*. A random forest classifier is trained to predict membrane pixels, using sparse manual labels as ground truth. If the automated segmentation is incorrect, the user supplies additional sparse labels where the most obvious errors are. Using this iterative approach, *ilastik* learns to classify pixels with a fraction of the labelling effort required for a fully automated approach. However, it is difficult to get a perfect segmentation from *ilastik*, and it appears to be primarily designed as an efficient approach to train a fully-automated pixel classifier. Vu and Manjunath (2008) and Straehle et al. (2011) also use an iterative sparse manual labelling approach, and both programs appear targeted towards an interactive semi-automated reconstruction. They use the sparse manual labels to adjust graph cut and watershed based segmentations respectively. As with *ilastik*, the user targets each iteration of labelling to the areas where the reconstruction is most incorrect. Interactive programs for semi-automated reconstruction also include contour-based methods, with at least two programs that use sparse user interaction to guide the propagation of active contours (Jeong, 2009; Jeong et al., 2010; Macke et al., 2008). Jones et al. (2013) take an interesting approach that does not use iterative labelling. Instead, the user is presented with a grid overlaid on the image, and asked to label the intersections of the grid with cell membranes. Their algorithm then finds paths between all labelled grid-membrane intersections. The sparse grid-based labelling divides the membrane labelling problem into many smaller problems, and

results in an improvement in performance over the fully-automated approach they benchmarked against. We also take a non-iterative approach to sparse labelling in this work, though in our case the sparseness is in the number of slices labelled rather than the number of pixels labelled in each slice.

Another approach to semi-automated reconstruction is to combine a relatively poor quality fully-automated volume reconstruction with a high quality manual centreline tracing for each neurite. In this approach, a *convolutional neural network* similar to that of Turaga et al. (2009) generates a dense volume segmentation, where every pixel is assigned to a reconstructed neurite segment. These automatically reconstructed segments are a significant *over-segmentation* of the true neurites and are not a sufficiently accurate reconstruction with which to perform neuroscience. However, each one is a reasonably accurate *local* reconstruction of a portion of true neurite. Independently, the centreline of each neurite in the volume is manually traced up to four times to ensure it is accurate. This centreline is then used to join the automatically generated segments together, threading them like beads on a string. The final result is an acceptably accurate volume reconstruction for the labelling effort of a skeleton reconstruction. This approach has been used for two recent studies that have generated new insights into patterns of connectivity in the retina (Briggman, Helmstaedter, and Denk, 2011; Helmstaedter et al., 2013). An interesting alternative approach to semi-automated reconstruction is that taken by Seung (2013). Using a similar *convolutional neural network* to generate a relatively poor quality fully-automated segmentation, the proof-reading of this segmentation is crowd-sourced by incorporating it into an online game. Previous efforts at "gamifying" science have been successfully applied to the problems of gene sequence alignment (Kawrykow et al., 2012) and protein structure prediction (Cooper et al., 2010).

The Helmstaedter et al. (2013) study reconstructed 950 neurons, and required over 20,000 hours of manual centreline tracing. It is therefore likely that this approach will not scale sufficiently well to be the final solution to the problem of generating reconstructions for large scale connectomics. To reconstruct a 1 mm^3 cortical column using such a semi-automated approach would take approximately 140 person years of manual tracing effort. While this is a substantial improvement on the $\sim$7,000 years a purely manual volume reconstruction would take, scaling this approach up to an entire mouse brain would require $\sim$70,000 person years of manual reconstruction effort just for the skeleton tracing.

3.5 Measures of segmentation accuracy

3.5.1 Binary pixel classification accuracy

A common approach to segmenting multiple neurites in electron microscope images is to attempt to identify the pixels representing the membrane that forms the boundaries between cells. If all the membrane pixels are identified correctly then the pixels representing each neurite interior will be isolated clusters of unlabelled pixels, separated by membrane pixels. These clusters can easily be identified using the *connected components* algorithm. Measures of *binary pixel classification accuracy* directly measure how well these membrane pixels are identified by a segmentation algorithm. A wide range of such measures have been proposed, and we shall discuss a selection of these in more detail below. All these measures can be calculated from a *binary confusion matrix*. Figure 3.1 shows three alternative representations of the binary confusion matrix. In the case of membrane detection, membrane pixels are labelled as 1 and non-membrane pixels are labelled as 0 in all three representations. In the first representation (3.1a), one of the segmentations is considered the *ground truth* segmentation and the other is considered the *reconstruction*, while in the other two representations (3.1b and 3.1c) neither segmentation necessarily has to be considered the ground truth. In all three representations we can classify pixels into one of four categories.

1. Pixels that are labeled membrane in both segmentations. These are the *true positives* (TP) in 3.1a and are referred to as a and n_{11} respectively in 3.1b and 3.1c.

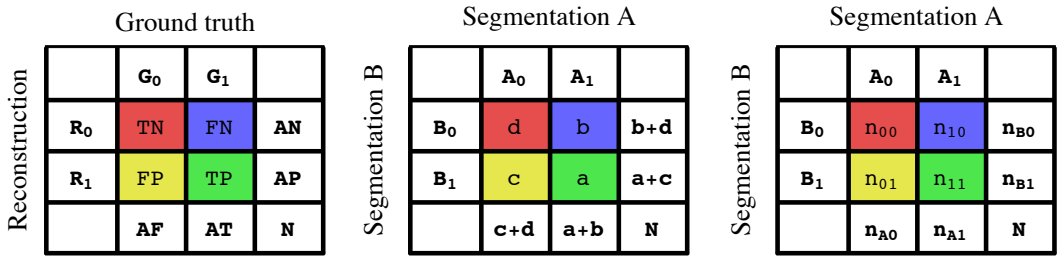
2. Pixels that are labeled non-membrane in both segmentations. These are the *true negatives* (TN) in 3.1a and are labelled d and n_{00} respectively in 3.1b and 3.1c.

3. Pixels that are labeled non-membrane in the first or ground truth segmentation and membrane in the second segmentation. These are the *false positives* (FP) in 3.1a and are labelled c and n_{01} respectively in 3.1b and 3.1c.

4. Pixels that are labeled membrane in the first or ground truth segmentation and non-membrane in the second segmentation. These are the *false negatives* (FN) in 3.1a and are labelled b and n_{10} respectively in 3.1b and 3.1c.

Jaccard index

Many measures of binary pixel classification accuracy were originally proposed for measuring ecological similarity. The earliest measure was proposed by Jaccard (1901a) to quantify the similarity between alpine regions in terms of the number of species they shared. It was defined as the number of species shared by two regions, divided by the total number of unique species



(a) True vs. positive notation. (b) a, b, c, d notation. (c) Object label notation

Figure 3.1: Three representations of the binary confusion matrix. The highlighted cells are used to calculate all the discussed similarity measures. The remaining cells can all be calculated from these. **(a)** Using the true vs. positive notation commonly used when one of the segmentations is considered the true segmentation (ground truth): TN = true negative; TP = true positive; FN = false negative; FP = False positive; AN = all negative; AP = all positive; AF = all false; AT = all true; N = total number of pixels. **(b)** Using the a, b, c, d notation commonly used in papers discussing binary similarity measures **(c)** Using the object label notation commonly used when extending such matrices beyond the binary labelling case: n_{ij} = the number of pixels in object i in segmentation A and in object j in segmentation B; n_{Xi} = the total number of pixels in object i in segmentation X. In the case of binary pixel classification there are only two objects, target (label 1) and background (label 0).

present across the two regions. In terms of membrane labelling accuracy, it is the number of pixels labelled as membrane in *both* segmentations divided by the number of pixels labelled as membrane in *either* segmentation ($\frac{n_{11}}{n_{11}+n_{10}+n_{01}}$). This measure was originally called the *coefficient de communauté* (*coefficient of community*) by Jaccard but is now commonly referred to as the *Jaccard index* (appendix B).

Pixel accuracy

The *Jaccard index* only considers *positive matches* in its computation. No credit is given for pixels that are labelled non-membrane in both segmentations. An alternative measure was proposed by Sokal and Michener (1958), which also considers *negative matches*. Introduced to quantify the similarity between species, it was defined as the number of attributes either *present* in both species or *absent* in both species, divided by the total number of attributes considered. In terms of membrane labelling accuracy, it is the fraction of total pixels that have the same label in both segmentations ($\frac{n_{11}+n_{00}}{n_{11}+n_{10}+n_{01}+n_{00}}$). Originally called the *simple matching co-efficient*, it is now commonly referred to as *pixel accuracy*.

A family of similarity measures

Both the *Jaccard index* and *pixel accuracy* can be considered as members of a wider family of similarity measures that differ only in whether they include *negative matches* (n_{00}) in their calculation and the relative weights assigned to *matches* (n_{11}, n_{00}) and *mismatches* (n_{10}, n_{01}). The group of measures which exclude negative measures was formally defined by Tversky (1977).

The *Jaccard index* is a member of this group, as is the popular *Dice co-efficient* (*coincidence index* in Dice, 1945). An equivalent group can also be defined for measures which include negative matches, with the addition of n_{00} to both the denominator and numerator. *Pixel accuracy* is a member of this second group. Table 3.1 lists several members of this family, although it can be infinitely extended by selecting different relative weights for *matches* and *mismatches*. All these measures are discussed in Lesot, Rifqi, and Benhadda (2009), who provide a good overview of this type of similarity measure. However, our grouping of measures based on equivalence to Tversky's measure splits their *type 2* measures into two groups. The measure defined by Russel and Rao (1940) is not considered an equivalent to Tversky's measure as it includes n_{00} in the denominator only, but is included here as it precedes all measures in the family except for the *Jaccard index*. For the Tversky equivalent measures in table 3.1, these relative weights are controlled by a single *relative mismatch weight* that sets the weight for both types of mismatch. However, in Tversky's original formulation, the weights for *false negatives* (n_{10}) and *false positives* (n_{01}) can be set independently. Thus the family of measures can be extended to cover cases where these two types of error have differing costs.

Precision and recall measures

The concepts of *precision* and *recall* are commonly associated with the problem of document search and retrieval. In this context *precision* is the fraction of *all retrieved* documents that are relevant to the search, while *recall* is the fraction of *all existing* relevant documents that are retrieved by the search. In terms of membrane labelling, *precision* is the number of pixels labelled as membrane in *both* segmentations divided by the total number of pixels labelled membrane in the *reconstruction* segmentation ($\frac{n_{11}}{n_{11}+n_{01}}$). *Recall* is the number of pixels labelled as membrane in *both* segmentations divided by the total number of pixels labelled membrane in the *ground truth* segmentation ($\frac{n_{11}}{n_{11}+n_{10}}$). Note that calculating *precision* and *recall* requires that one of the segmentations is considered as the *ground truth*. However, *precision* and *recall* are usually combined to generate a single composite measure that does not depend on which segmentation is considered the *ground truth*. The most common method of combining *precision* and *recall* is to take their *harmonic mean*, popularly known as the *f-measure*. It is commonly used as a measure of membrane segmentation accuracy, as it is not strongly affected by the large imbalance between the numbers of membrane and non-membrane pixels. Other methods of combining *precision* and *recall* have been proposed, including the *arithmetic mean* (Kulczynski, 1927) and the *geometric mean* (Ochiai, 1957). *Precision* and *recall* were introduced in the context of document retrieval by Kent et al. (1955), with *precision* referred to as the *pertinancy factor*. However, exactly equivalent measures were discussed ten years earlier by Dice (1945)

Reference	Relative mismatch weight	Binary definition
Variants of Tversky's similarity measure excluding negative matches		
Jaccard (1901a)	1	$S = \frac{n_{11}}{n_{11}+n_{01}+n_{10}}$
Dice (1945)	$\frac{1}{2}$	$S = \frac{2n_{11}}{2n_{11}+n_{01}+n_{10}}$
Sorenson (1948)	$\frac{1}{4}$	$S = \frac{4n_{11}}{4n_{11}+n_{01}+n_{10}}$
Anderberg (1973)	$\frac{1}{8}$	$S = \frac{8n_{11}}{8n_{11}+n_{01}+n_{10}}$
Sokal and Sneath (1973)	2	$S = \frac{n_{11}}{n_{11}+2n_{01}+2n_{10}}$
Tversky equivalents including negative matches		
Sokal and Michener (1958)	1	$S = \frac{n_{11}+n_{00}}{n_{11}+n_{01}+n_{10}+n_{00}}$
Sokal and Sneath (1963)	$\frac{1}{2}$	$S = \frac{n_{11}+n_{00}}{n_{11}+\frac{1}{2}(n_{01}+n_{10})+n_{00}}$
Rogers and Tanimoto (1960)	2	$S = \frac{n_{11}+n_{00}}{n_{11}+2(n_{01}+n_{10})+n_{00}}$
Non-Tversky measures including negative matches		
Russel and Rao (1940)	n/a	$S = \frac{n_{11}}{n_{11}+n_{01}+n_{10}+n_{00}}$

Table 3.1: Binary similarity measures. There are a range of binary similarity measures that differ only in whether they include *negative matches* in their calculation and the relative weight they assign to *matches* (n_{11}, n_{00}) and *mismatches* (n_{10}, n_{01}). The group of measures which exclude negative measures was formally defined by Tversky (1977). An equivalent group can also be defined for measures which include negative matches, with the addition of n_{00} to both the denominator and numerator. The measure defined by Russel and Rao (1940) is not considered an equivalent to Tversky's measures as it includes n_{00} in the denominator only.

in the context of ecology, where they are referred to as *association indexes* and do not require the concept of a *ground truth*. Dice combines his *association indexes* into a single *coincidence index*, which is exactly equivalent to the *f-measure* harmonic mean of *precison* and *recall*. The earliest discussion of these concepts appears to be by Kulczynski (1927) where *precision* and *recall* are combined by taking their arithmetic mean. *Precision*, *recall* and the various methods of combining them into a single composite measure are formally defined in table 3.2.

Name	**Reference(s)**	**Binary definition**
Precision	Dice (1945); Kent et al. (1955); Kulczynski (1927)	$Pr = \frac{n_{11}}{n_{11}+n_{01}}$
Recall	Dice (1945); Kent et al. (1955); Kulczynski (1927)	$Re = \frac{n_{11}}{n_{11}+n_{10}}$
Arithmetic mean	Kulczynski (1927)	$S = \frac{Pr+Re}{2} = \frac{1}{2}(\frac{n_{11}}{n_{11}+n_{01}} + \frac{n_{11}}{n_{11}+n_{10}})$
Geometric mean	Ochiai (1957)	$S = \sqrt{Pr \times Re} = \frac{n_{11}}{\sqrt{n_{11}+n_{01}}\sqrt{n_{11}+n_{10}}}$
Harmonic mean	Dice (1945)	$S = \frac{2(Pr \times Re)}{Pr+Re} = \frac{2n_{11}}{2n_{11}+n_{01}+n_{10}}$

Table 3.2: Precision and recall based measures. *Precision* and *recall* are commonly associated with document retrieval, in which context they were described by Kent et al. (1955). However, these concepts were discussed much earlier by Kulczynski (1927) and Dice (1945). Various methods of combining *precision* and *recall* into a single composite measure have been proposed. The most common approach is to take their *harmonic mean*, which is known as the *f-measure*. It is exactly equivalent to the *Dice index* and a variant of Tversky's measure (table 3.1).

Weaknesses of binary pixel classification measures

We have discussed a family of *binary pixel classification* similarity measures. These are all based on a binary labelling, where each pixel can have a value of either 1 or 0. When segmenting electron microscope images of neurons, this binary labelling is almost always a *membrane labelling*, where membrane pixels are labelled with 1 and non-membrane pixels are labelled with 0. One of the key issues with using a similarity measure based on a binary labelling is that the cost of mislabelling any single pixel is the same. However, some pixels are more important than others when it comes to the accuracy of the final segmentation. The simplest method of converting a binary membrane labelling into a segmentation of neurite cross-sections is to use the *connected components* algorithm. If all the membrane pixels are identified correctly then the pixels representing each neurite interior will be isolated clusters of non-membrane pixels. If a few interior pixels at the edge of one of these isolated clusters are incorrectly labelled as

membrane, the consequence will be a small error in the shape and size of the reconstructed cross-section. However, there will be no topological error. All true neurite cross-sections will still be represented in the segmentation by an isolated cluster of pixels, and the recovered connectivity between neurons will be accurate. Mislabelling membrane pixels as non-membrane can be more serious, as the correct segmentation of neurite cross-sections is dependent on isolating these clusters by correctly classifying thin bands of membrane pixels. If a few pixels comprising the membrane separating two isolated clusters are incorrectly mislabelled as non-membrane, then these two clusters could end up joined by a "bridge" of incorrectly labelled pixels. Therefore, the two neurite cross-sections they represent will be incorrectly merged. This will result in a serious error in the topology of the reconstruction, resulting in incorrect connectivity between neurons. It is also possible for a single neurite cross-section to be incorrectly split into two clusters if a line of pixels running through it are mislabelled as membrane.

In order to capture the different costs of misclassifying various pixels, a similarity measure must therefore consider how well pixels are grouped together in the final segmentation. There are several possible approaches to do this. In section 3.5.2 we discuss two that have been used to assess segmentation similarity when reconstructing neurites from electron microscope images. In section 3.5.3 we describe our preferred approach for addressing this issue.

3.5.2 High level similarity measures

While many studies evaluating automated reconstruction methods report measures of binary pixel classification accuracy, some studies report higher level similarity measures. The most common of these is the *Rand index* (Rand, 1971). This measure considers all possible pairs of pixels within each segmentation, and allocates them to one of four classes. If the segmentations are labelled A and B, these four classes contain:

1. Pairs of pixels that are in the *same* object in A and the *same* object in B
2. Pairs of pixels that are in *different* objects in A and the *same* object in B
3. Pairs of pixels that are in the *same* object in A and *different* objects in B
4. Pairs of pixels that are in *different* objects in A and *different* objects in B

The *Rand index* is calculated by dividing the number of pixel pairs in classes (1) and (4) by the total number of pixel pairs. Turaga et al. (2009) have also determined how to directly optimise this higher level measure using a *convolutional neural network* (CNN). The same group has also developed an alternative high level similarity measure which they have also managed to optimise using a CNN. This measure is called the *warping error* (Jain et al., 2010).

When calculating this measure, pixels in the ground truth binary labelling can be "flipped" from zero to one and vice-versa so long as no pixel clusters are split or merged. This permits a *warping* of the ground truth labelling to more closely match the algorithm-generated labelling, while ensuring that no changes are made to the topology of the ground truth segmentation. The *warping error* is simply the binary pixel error between the algorithm-generated segmentation and the most similar warping of the ground truth. In calculating the *warping error*, a decision must be made regarding the maximum amount of warping to permit. Most studies that use this measure appear to permit unlimited topology preserving warping. However, this will shrink any unmatched algorithm-generated objects to a single pixel, which is probably not representative of the desired penalty for unmatched objects in most reconstruction scenarios. This measure is also computationally intensive to calculate.

3.5.3 Overlap as a measure of similarity

At the end of section 3.5.1, we concluded that a similarity measure should consider how well pixels are grouped together in the final segmentation. One way to do this is to consider the similarity of corresponding objects in the two segmentations. However, in order to do this a correspondence between the labels in the two segmentations must be established. To establish this correspondence, we consider the *overlap* between all possible pairings of objects between the two segmentations. The overlap between two objects is the *intersection* of the two objects divided by their *union*. If the objects considered are a pair of objects from two segmentations, then the *intersection* of the objects is the number of pixels that are members of *both* objects and the *union* of the objects is the number of pixels that are members of *either* object. It can be demonstrated that the *overlap* between each pair of objects is equivalent to their *Jaccard index* (see figure 3.2).

In order to determine the similarity between an algorithm-generated segmentation and a ground truth segmentation, we first calculate the *overlap* for all possible pairings of objects between the two segmentations. This forms the list of *candidate pairings*. Next, the candidate pairing with the highest overlap is identified and moved to a list of *matched pairings*. All remaining candidate pairings containing one or other member of the matched pair are removed from the candidate list. This process is repeated, starting with the identification of the pairing with the highest overlap from the current candidate list. This is continued until no candidate pairings with non-zero overlap remain.

At this point, we have established a one-to-one correspondence between objects in the two segmentations. In order to measure the similarity between the objects in each matched pair, we can use any of the binary similarity measures discussed in section 3.5.1. We choose to use the

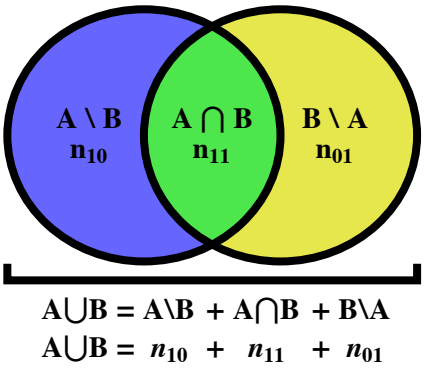


Figure 3.2: Overlap is defined in terms of the *intersection* ($A \cap$B) and *union* ($A \cup$B) of two objects. Translated into the terms used in section 3.5.1 to discuss the *binary pixel classification accuracy* between two segmentations: the *intersection* of two segmentations is the number of pixels given the same label in both segmentations (n_{11}) and the *union* is the total number of pixels given that label in *either* segmentation ($n_{10} + n_{11} + n_{01}$). It can be seen that, for a matched pair of labels representing the same object, *overlap* $= \frac{intersection}{union} = \frac{n_{11}}{n_{10}+n_{11}+n_{01}} =$ *Jaccard index*.

same overlap measure used to establish the correspondence between objects in the two segmentations. This is identical to the binary *Jaccard index* between each pair of objects. To generate a global measure of similarity between the two segmentations, we must combine the overlaps for these matched pairings in a sensible manner. To do this, we calculate the total *matched overlap* by summing the overlap across all matched pairings. We then make use of the concepts of *precision* and *recall* introduced in section 3.5.1. *Precision* is defined as the mean matched overlap for algorithm-generated objects, and is calculated by $\frac{matched\ overlap}{number\ of\ algorithm\ generated\ objects}$. *Recall* is defined as the mean matched overlap for ground truth objects, and is calculated by $\frac{matched\ overlap}{number\ of\ ground\ truth\ objects}$. We then take the harmonic mean of precision and recall to generate the *overlap f-measure*. As discussed in section 3.5.1, this a commonly used approach for combining precision and recall measures. The process of calculating the *overlap f-measure* for two segmentations is described in detail in algorithm 2 (section 5.3.2).

Polak, Zhang, and Pi (2009) suggested a very similar measure based on summing the overlap between pairs of objects in two segmentations. However, there are some key differences. Firstly, a one-to-one pairing between objects in the two segmentations is not enforced. Instead, when an algorithm-generated object intersects with multiple ground truth objects, a weighted average of its overlap with all intersecting ground truth objects is taken. We would suggest our one-to-one matching approach is more intuitive, although it may be that there is little difference in the total matched overlap calculated using the two approaches. Secondly, the overlap for each pairing is weighted by the size of the ground truth object. This gives more credit for

matching large ground truth objects than for matching small ones. We would argue that this is not a desired property for assessing the accuracy of neural circuit reconstructions, where it is extremely important to successfully track neurites when they become small. Finally, their measure is directional, assessing reconstruction accuracy from the perspective of the ground truth. This makes it a measure of *recall* only, with no penalty for algorithm-generated objects that do not intersect with *any* ground truth object. We would argue that a symmetric measure such as ours, that considers both *precision* and *recall*, is more appropriate for assessing reconstruction accuracy.

The *Rand index* discussed in section 3.5.2 avoids the requirement to establish label correspondence as it considers the consistency of labelling between pairs of pixels across the two segmentations. Therefore, it could be argued that it is a more straightforward measure of segmentation similarity. However, for our work we are able to make use of overlap as the target output for one stage of our reconstruction algorithm. Therefore it is natural to also use it as the basis of our reconstruction similarity measure.

3.5.4 Measures of 3D reconstruction accuracy

We evaluate 3D reconstruction accuracy using a pair of measures. The *matched segment run-length* captures the length of successfully reconstructed fibre segments, while the *matched f-measure* captures the overall proportion of fibre cross-sections that are well found. We introduce these measures in section 7.2, where we also discuss alternative measures used to assess 3D neural circuit reconstructions in other studies.

Chapter 4

Data collection and curation

In this chapter we describe the acquisition of our image data and the generation of the corresponding manually labelled ground truth. We discuss the quality of the manual labelling and the publication of our data set as a community resource.

4.1 Image acquisition

A sample from the molecular layer of a perfusion-fixed mouse cerebellum was classically stained using a reduced OTO method (Willingham and Rutherford, 1984). The sample was then imaged in sagittal sections using an NVision 40 Focussed Ion Beam Scanning Electron Microscope (FIBSEM), at an isotropic resolution of 9.3 nm. The acquired images were registered into a common reference frame using the *Linear Stack Alignment with SIFT* plug-in for Fiji (SIFT: Lowe, 2004; Fiji: Schindelin et al., 2012). A 2548×852×512 voxel sub-volume (Häusser lab ID: *OReilly::block03*) was used for this work. Details of the SIFT alignment parameters and sub-volume offsets are provided in appendix A. Sample preparation and image acquisition were performed by Sarah Rieubland, Arnd Roth and Arifa Naeem of the Häusser Lab at the UCL Wolfson Institute for Biomedical Research. Fuller details of sample preparation and image acquisition protocols have kindly been provided by Sarah Rieubland in appendix A.

4.2 Generating ground truth labels

All ground truth labelling was done using the TrakEM2 plug-in for Fiji (Cardona et al., 2009).

4.2.1 Initial 2D labelling

Four full slices of the *block03* data set were carefully labelled by the author (slices 1, 171, 341, 511). Membrane was labelled as a single TrakEM2 AreaList object. In each slice, all external membrane was traced using a 3 pixel wide brush. This left only cell interiors unlabelled. All cell interiors were then labelled as either *normal fibre*, *fibre bouton*, *closed dendritic spine cross-section* or *other*. After this process, all pixels were labelled. Each slice was then split into

two. These half-slice ground truth data sets were given the identifiers 1-4 (left half: 1, 171, 341, 511) and 5-8 (right half: 1, 171, 341, 511). To make the *membrane* ground truth, the membrane labelling was used with no further processing. To make the sparse *fibre-only* cross-section ground truth, all the cell interior labels except *other* were merged and individual cell cross-sections identified via connected components analysis in Matlab. To make the dense *all-cell* cross-section ground truth, *all* the cell interior labels were merged and individual cell cross-sections identified via connected components analysis in Matlab. All connected component cross-sections in the *fibre-only* and *all-cell* ground truth data sets were dilated by 2 pixels to account for the width of the membrane labelling brush. These three data sets were used for training and evaluating all 2D algorithms discussed in chapter 6. For our algorithm, only the *fibre-only* ground truth was used. For training and evaluating ilastik, all three ground truth data sets were used. Figure 4.1a shows example manual labelling for data set 8. The *fibre-only* ground truth consists of all green cross-sections individually dilated by 2 pixels. The *all-cell* ground truth consists of all green and red cross-sections individually dilated by 2 pixels.

4.2.2 Full 2D membrane labelling

For a 1274×852×151 sub-volume of the *block03* data set, all external cellular membrane was traced using a 3 pixel brush, with membrane labelled as a single TrakEM2 AreaList object. This sub-volume comprised the right-hand side of all slices from 161 to 311 inclusive. Outside this sub-volume, membrane was labelled across the full extent of every fifth slice (i.e. all slice numbers ending in 1 or 5). For computational reasons each slice was split into overlapping left and right sections for tracing. These overlapping sections were then merged prior to the 3D labelling process (section 4.2.3). This 2D membrane labelling was performed by Sophie Gordon-Smith, Rashmi Gamage, Trisha Patel, Kylie Wong and Maja Boznakova. The 2D labelling process was initially overseen by the author and Sarah Rieubland. Maja Boznakova supervised the later stages, and Arnd Roth provided valuable support throughout the process.

4.2.3 3D neurite labelling

The 2D membrane labelling described in section 4.2.2 provided the basis for labelling individual neurites in 3D. A new TrakEM2 AreaList object was created for every 3D segment of axon or dendrite present in the *block03* data set. These included segments of parallel fibres, interneuron axons, interneuron dendrites, Purkinje cell dendrites and some possible climbing fibres. Each individual neurite segment was carefully followed throughout the volume, and all cross-sections belonging to it were labelled as the same TrakEM2 object via selective flood filling of the 2D membrane labelling. Glial cells were not labelled in 3D, and all unlabelled pixels are putative

glia interior. Each reconstructed neurite segment was examined in 3D and further labelled as either *fibre* or *other* based on its morphology and geometry. The *fibre* class includes all parallel fibres, but also any other axons that run at within $\sim$45° of perpendicular to the imaging plane. This 3D neurite labelling was performed by Maja Boznakova and Rashmi Gamage. Maja Boznakova supervised the 3D labelling work, with final proof-reading performed by the author.

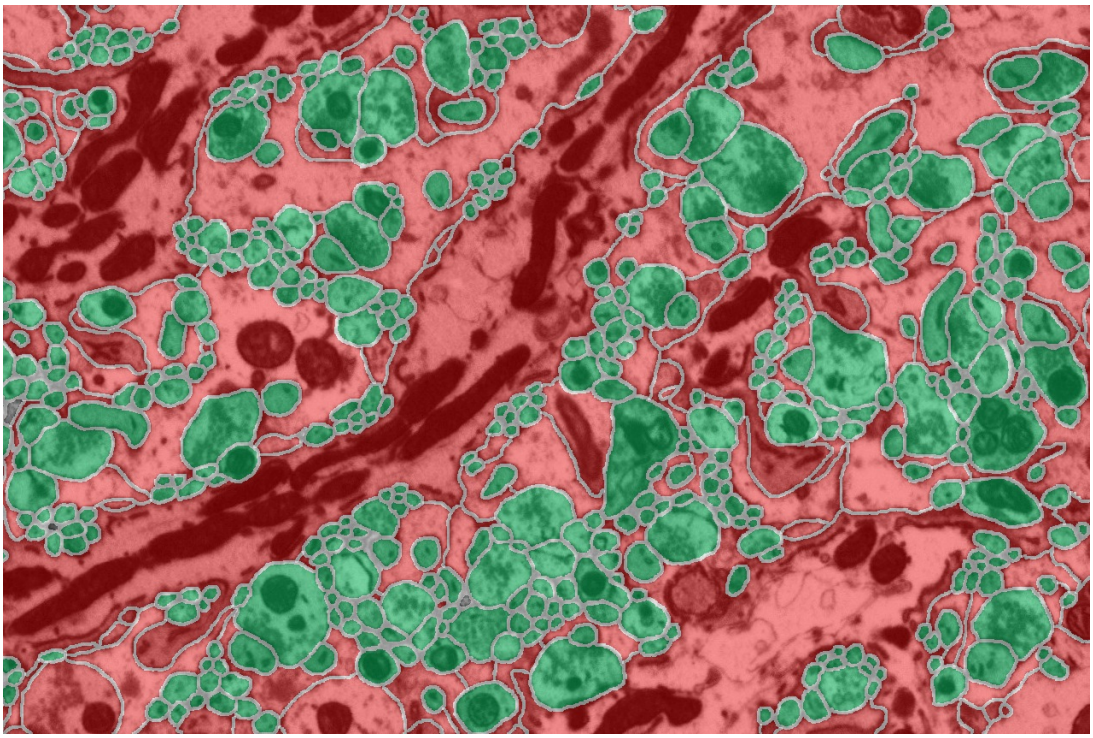
The final 2D membrane labelling and 3D neurite labelling both span the entire extent of each slice with no discontinuities in labelling. However, we split the final labelled volume into non-overlapping sub-volumes for use in this work. Each slice was split down the middle, with the left side of each slice given an identifier of the form $1nnn$ and the right hand of each slice given an identifier of the form $2nnn$. In both cases, nnn is the slice number padded with leading zeros (e.g. 1001 for the left hand side of slice 1 and 2311 for the right hand side of slice 311). For the evaluation of our 3D reconstruction algorithm in 7 we use a 151-slice data set comprising slices 2161-2311 inclusive. Figure 4.1b shows example manual labelling for slice 2511 (corresponding to data set 8 in the initial 2D labelling). The 3D *fibre-only* ground truth used to develop and evaluate the 3D algorithm consists of all objects classified as *fibre*, with their 2D cross-sections individually dilated by 2 pixels.

4.2.4 Data quality

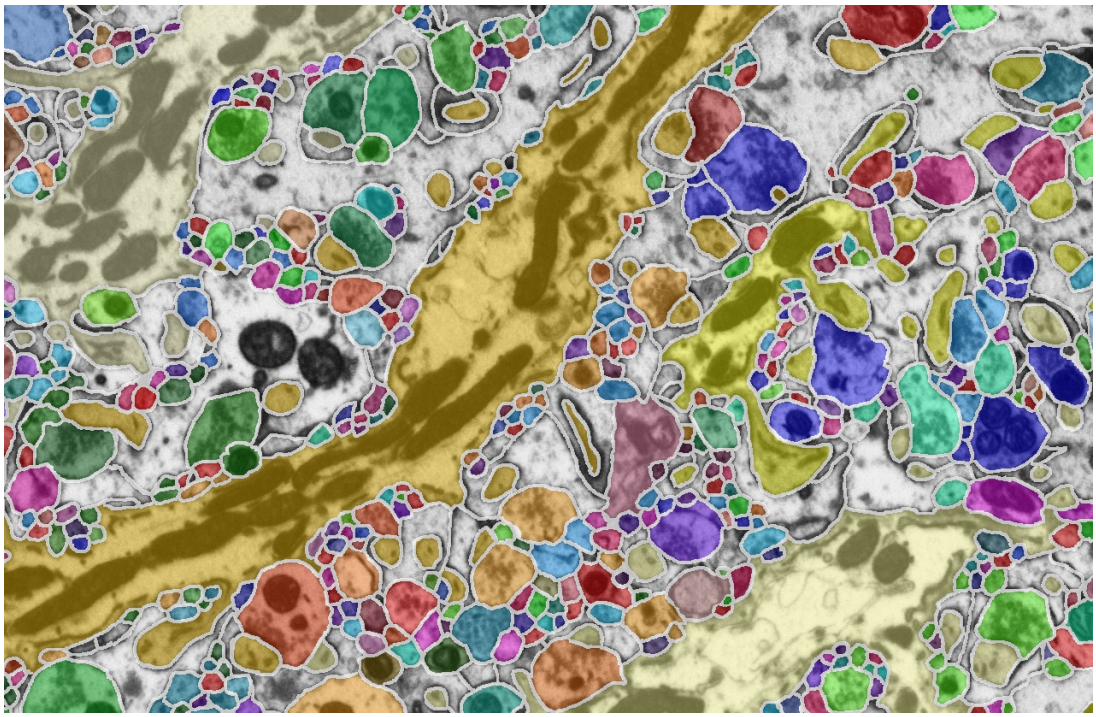
The quality of the full 2D membrane labelling was ensured by training tracers and regularly reviewing their work. Each tracer spent a week tracing a set of images with known labelling, with ready access to scientists experienced at interpreting the electron microscope images. Their work was regularly reviewed during the training process and they were only permitted to start tracing unlabelled images when the quality of their tracing was satisfactory. The quality of their labelling was regularly reviewed throughout the period they were tracing. Nonetheless, there were some issues with the quality of the raw 2D membrane labelling.

Variance in membrane positioning

Tracers were asked to strike a balance between speed and accuracy, so there was some discrepancy in the exact position of the membrane between the raw 2D labelling and the underlying image data in places. For all singly traced regions, these differences did not affect the topology of the reconstruction. However, for the overlapping region common to the left and right hand sections of each slice, these positioning discrepancies caused some problems that needed to be resolved. Combining the membrane labels from the left and right sections resulted in very thick membrane in places, along with some holes where the left and right membrane labels did not fully overlap. The combined membrane labelling in this central overlapping region needed to



(a) 2D membrane (white) / fibre (green) / non-fibre (red) labelling.



(b) 3D individual object labelling.

Figure 4.1: Examples of **(a)** Membrane (white) / fibre (green) / non-fibre (red) labelling used for 2D algorithm development and evaluation. The 2D *fibre-only* ground truth consists of all green cross-sections individually dilated by 2 pixels. The 2D *all-cell* ground truth consists of all green and red cross-sections individually dilated by 2 pixels. **(b)** Individual object labelling used for 3D algorithm development and evaluation. Unlabelled objects are cross-sections of non-neuron support cells (glia). The 3D *fibre-only* ground truth used to develop and evaluate the 3D algorithm consists of all objects classified as *fibre*, with their 2D cross-sections individually dilated by 2 pixels. Labelling shown for right half of slice 511 (2D data set 8; 3D slice 2511).

be extensively edited to ensure a similar level of quality to the labelling in the singly traced regions. This editing was performed prior to the 3D labelling process and no such problems remain in the final ground truth data set.

Topological errors in membrane labelling

Tracers primarily used 2D "within slice" image information to place membrane labels. As a result, there were some topological errors in the raw membrane labelling where the membrane evidence provided by a single image was difficult to interpret. For all cells reconstructed in 3D, any topological errors introduced by errors in membrane labelling were corrected during the 3D labelling process. However, glial cells were not reconstructed in 3D and therefore some topological errors in the glial membrane tracing are likely to remain. For any analysis using the final ground truth data set that requires accurate glial topology, further proof-reading and correction of glial membrane labelling will be required.

Inter-person variability in membrane tracing

As the author had also labelled some of the slices traced during the full 2D labelling, an opportunity arose to evaluate the *inter-person variability* in membrane tracing. The 2D membrane labelling was converted into a cell interior labelling by finding the *connected components* of the non-membrane pixels. The overlap between the cross-sections in each pair of labellings was calculated and is presented in table 4.1. The median overlap is 0.7, although the overlap for slice 341 seems to be lower than for the other two slices. This overlap is high but not perfect, and may be a useful accuracy target for considering whether an automated 2D reconstruction is acceptable. However, it is possible that these overlap calculations might still incorporate some effect of uncorrected glial membrane labelling. Efforts were made to exclude these effects in the *edited* overlap by only including cross-sections from each labelling where they have an overlap >0.1 with a cross-section in the final 3D *fibre* ground truth. This threshold was chosen as it resulted in the number of cross-sections considered in each labelling to be within ± 1 of the number of cross-sections in the 3D *fibre* ground truth. This should result in a near optimal overlap f-measure. However, further analysis is required before this inter-person overlap could be considered as a threshold for determining whether an automated reconstruction is acceptable.

4.3 Publication of a "gold standard" reference data set

The publication of "gold standard" benchmark data sets in other fields of computer vision have resulted in an increase in the number of researchers applying their machine learning techniques to these data sets. Until recently no such data set has existed for the problem of reconstructing neurons from electron microscope images. However, over the past two years, two data sets

Tracers	Slice	Section	Overlap f-measure	
			Raw	Edited
MOR vs RG	1	Left	0.69	0.70
MOR vs SBS	341	Left	0.58	0.60
MOR vs SBS	511	Left	0.69	0.70
		Mean	**0.64**	**0.65**
		Median	**0.69**	**0.70**

Table 4.1: Inter-person labelling variability. The *raw* overlap is the overlap for all cell interiors. The *edited* overlap is the overlap for cell interiors that have an overlap >0.1 with a cross-section in the final 3D *fibre* ground truth. This threshold was chosen as it resulted in the number of cross-sections considered in each labelling to be within ± 1 of the number of cross-sections in the *fibre* 3D ground truth.

have been released as part of a challenge workshop at the annual International Symposium on Biomedical Imaging (ISBI). The first is a Drosophila ventral nerve cord data set with 2D binary labelling (Cardona, 2012), and the second a mouse cortex data set with 3D object identity labelling (Shaar, 2013). These are discussed further in section 6.6.2. These new benchmarking data sets are welcome contributions to the community, and we plan to add to this growing community resource by publishing the *block03* data set used in this work. This data set covers a significantly greater volume than the ISBI data sets, and we will publish both the electron microscope images and the ground truth labelling for this data set in the open access Cell Centered Database (CCDB). The contribution of another high quality reference data set to the field will hopefully facilitate an expansion in the number of researchers working on the challenging problem of reconstructing neurons from electron microscope images. We hope that it will encourage researchers without access to neuroscience collaborators and an electron microscope to apply their machine learning techniques to this problem. Hopefully this will accelerate progress towards the reconstruction and analysis of usefully large micro-circuits. In particular, the highly ordered structure of the parallel fibres in our data set will enable researchers new to the field to start by applying less computationally intensive 2+1D processing techniques. While a similar set of electron microscope images for a larger volume of the mouse cerebellum have already been published (Bushong and Deerinck, 2013), the corresponding ground truth labelling is much smaller than ours and it has not yet been published.

Chapter 5

Modelling neural fibres

In our model-based method for reconstructing parallel fibres we represent fibre cross-sections with circles. In this chapter we justify this representation both theoretically and empirically. We then demonstrate that predicting the *overlap* of candidate circles with ground truth fibres is sufficient to generate a high quality reconstruction. Finally, we introduce our use of annular histograms of Basic Image Features (BIFs) as "fibreness" feature vectors to assess the image evidence for candidate circles.

5.1 Overview

In sections 5.2 and 5.3 we demonstrate that circles are a reasonable representation of fibre cross-sections. To do this we:

1. Introduce the concept of *overlap* and derive theoretically justified thresholds for *high*, *medium* and *low* overlap based on overlapping circles.

2. Demonstrate that for the vast majority of fibre cross-sections the best circle representation has *high overlap* with the true polygon representation.

3. Demonstrate that our overlap thresholds are applicable to fibre cross-sections by showing that the threshold above which neuroscientists consider the circle representation of a fibre cross-section acceptable is similar to our *high overlap* threshold.

4. Demonstrate that a circle representation of fibre cross-sections is suitable for deriving models of neural networks at various levels of abstraction by comparing key properties of these models when constructed using true fibre cross-sections and circle approximations.

5. Introduce the concept of a *ground truth overlap volume* to simultaneously represent how well each of a large set of candidate circles represents any of the fibre cross-sections present in an image. We demonstrate that accurately predicting this *ground truth overlap*

volume for an image is sufficient for finding a set of circles that well represents the fibre cross-sections present in it.

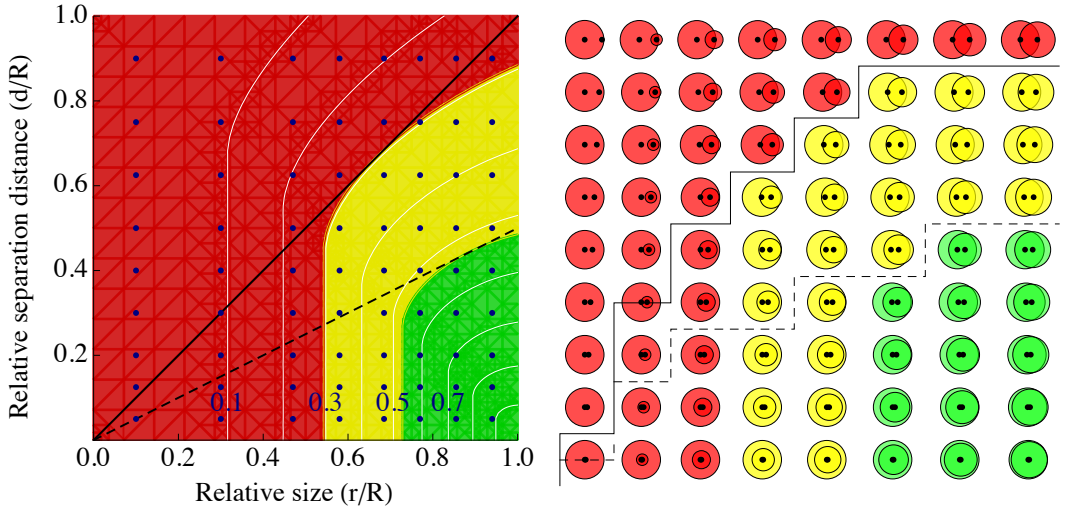
In section 5.4 we introduce Basic Image Features (BIFs) and our use of annular BIF histograms to assess the image evidence for candidate circles. In chapter 6 we will use these annular BIF histograms as "fibreness" feature vectors to predict the *ground truth overlap volume* for an image.

5.2 Circles as a representation of fibre cross-sections

5.2.1 Circles have high overlap with fibre cross-sections

In this work we have selected *overlap* as our chosen measure of segmentation similarity (see section 3.5.3). Therefore, if we are to claim that circles are an appropriate representation of fibre cross-section, we must show that they can have high overlap with the cross-sections they represent. This raises the question of what degree of overlap should be regarded as "high". We have derived theoretical thresholds for *low*, *medium* and *high* overlap on the basis of the mutual *threading* of pairs of circles. We define two circles as *threaded* if they each contain the other's centre within their radius and as *well-threaded* if they each contain the other's centre within half their radius. We then define *low*, *medium* and *high* overlap in terms of threadedness.

We can visualise both *threadedness* and *overlap* for all possible pairs of circles by plotting these properties as a function of the *relative size* and *relative separation distance* of circle pairs. Relative size is the radius of the smaller circle (r) normalised by the radius of the larger circle (R). Relative separation distance is the distance between the two circle centres (d) normalised by the radius of the larger circle (R). This visualisation can be seen in figure 5.1a. We define the boundary between *low* and *medium* overlap as the maximum overlap achievable by an *unthreaded* pair of circles (0.296) and the boundary between *medium* and *high* overlap as the maximum overlap achievable by a pair of circles that are not *well-threaded* (0.532). Figure 5.1b shows example circle pairs corresponding to a sampling of the parameter space in 5.1a. Examining these example circle pairs, we would argue that our selection of *medium* and *high* overlap thresholds results in a reasonable partition of these circle pairs into *high overlap* (green), *medium overlap* (yellow) and *low overlap* (red) examples. We would further argue that our *medium* and *high* overlap thresholds result in a better partition than the *threaded* and *well-threaded* thresholds they are derived from (solid and dashed lines respectively). In particular, we would argue that the overlap-based partition is more appropriate than the threading-based partition when a pair of circles are well-aligned but mismatched in size (lower left portion of figure 5.1b).



(a) Overlap as a function of relative size and separation distance

(b) Pairs of circles corresponding to the parameters marked by blue dots in (a)

Figure 5.1: **(a)** The overlap between pairs of circles as a function of their relative size (smaller radius over larger radius) and separation distance (expressed as a fraction of the larger radius). White contours are iso-overlap lines, with corresponding overlap values indicated in blue. Red, yellow and green zones indicate areas of *low*, *medium* and *high* overlap respectively. Black solid and dashed lines separate the *unthreaded* (above the solid line), *threaded* (below the solid line) and *well-threaded* (below the dashed line) areas of the parameter space. A *threaded* pair of circles each contain the other's centre within their radius. A *well-threaded* pair of circles each contain the other's centre within half their radius. The boundary between *low* and *medium* overlap is set to the maximum overlap achievable by an *unthreaded* pair of circles (0.296). The boundary between *medium* and *high* overlap is set to the maximum overlap achievable by a pair of circles that are not *well-threaded* (0.532). The blue dots indicate the parameters for the corresponding circle pairs shown in (b). **(b)** Pairs of circles corresponding to the parameters marked by blue dots in (a). Colouring indicates *low* (red), *medium* (yellow) and *high* (green) overlap. The solid and dashed lines separate the *unthreaded* (above the solid line), *threaded* (below the solid line) and *well-threaded* (below the dashed line) pairs.

Having derived thresholds for *low*, *medium* and *high* overlap, we must now demonstrate that circles have high overlap with the fibre cross-sections they represent. Figure 5.2 shows the histogram of overlaps between 12,610 fibre cross-section polygons and their corresponding maximum overlap circles. It can be seen that the vast majority (98.8%) of fibre cross-sections have high overlap with their best fitting circles and none have low overlap (the minimum polygon-circle overlap is 0.364). In terms of our theoretically derived overlap thresholds, circles are therefore a suitable representation of fibre cross-sections.

Finally, given that our theoretical overlap thresholds were derived by considering circle pairs, we must validate that it is reasonable to apply them to polygon-circle pairs. In figure 5.3 we show a random sample of fibre polygon-circle pairs exhibiting a range of overlaps. A similar sample was shown to ten neuroscience colleagues, who were asked to make a qualitative judgement on the range of overlaps for which circles represent an *acceptable fit* for the underlying fibre polygons. The mean judgement was that circles with an overlap of 0.6 and above represent an *acceptable fit*. While this threshold is somewhat higher than our theoretically derived *high overlap* threshold of 0.532, we would argue that the two thresholds are sufficiently close that it is reasonable to apply the theoretically derived overlap thresholds to polygon-circle pairs. Even if it were not, the fraction of fibre polygon-circle pairs exceeding the somewhat stricter human threshold is only marginally lower at 96.1% (see figure 5.2). Therefore the conclusion that circles have sufficiently high overlap with fibre cross-section polygons to be a reasonable representation of them would still stand.

5.2.2 Circles are a suitable abstraction for modelling neural fibres

We have demonstrated that circles are a suitable representation of fibre cross-sections in terms of our chosen segmentation similarity measure. However, in order to conclusively demonstrate that circles are a suitable representation of fibre cross-sections we must consider the possible uses to which a reconstruction of a neural network will be put. Below we discuss the common levels of abstraction used for the analysis of neural circuits and evaluate the suitability of a circle representation for each of these. We will show that for most levels of abstraction a circle representation is likely to be as suitable as one that uses polygonal cross-sections.

Known limitations

This analysis is based on a region of the brain where the majority of neural fibres run approximately perpendicular to the image plane. Further analysis is recommended if a circle-based reconstruction is being considered for a brain region with more varied geometry. This analysis can be repeated for such a region by considering the full 3D ground truth when extracting fi-

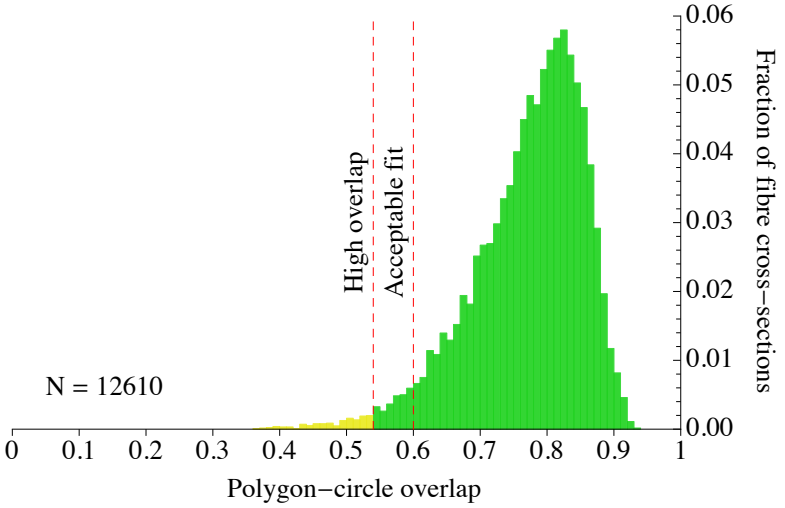


Figure 5.2: Distribution of the overlap between polygons representing fibre cross-sections and their *maximum overlap* circles (right axis). Bar colours indicate *low* (red), *medium* (yellow) and *high* (green) overlap. 98.8% of fibre cross-sections have high overlap with their best-fitting circles and none have low overlap. **Dashed red lines:** Theoretical *high overlap* and mean human *acceptable fit* thresholds. 96.1% of fibres exceed the human threshold of 0.6.

bre cross-sections. Additionally, when considering the suitability of a circle representation for various representations of neural circuits, we only consider its suitability as an approximation of fibre cross-section. There are secondary considerations regarding the impact of a circle approximation on the ability to correctly attribute synapses to neurons and to successfully track neurite branching. Synapses form the chemical connections between neurons and incorrectly identifying which neurons are associated with a given synapse will result in the wrong pair of neurons being connected. Correctly tracking branches is also critical. Failing to correctly associate a branch with its parent neuron can cause a large number of synapses to be attributed to the wrong neuron, resulting in large errors in connectivity. We discuss these issues in more detail below. However, in this work we restrict ourselves to the problem of reconstructing neural cross-sections and do not consider the issues of synapse attribution and branching.

Synapse attribution: Until recently any issues regarding synapse attribution would have been largely moot, as even recent studies still overwhelmingly use manual synapse reconstruction (Merchan-Perrez et al., 2009; Lu, Esquivel, and Bower, 2009; Chklovskii, Vitaladevuni, and Scheffer, 2010; Mishchenko et al., 2010). When synapses are being manually reconstructed, the manual assignment of synapses to neurons is no more work when using a circle representation than when using polygons. However automated approaches for synapse identification are now being proposed (Kreshuk et al., 2011) and even claim a lower error rate than a single human. The Kreshuk et al. technique labels post-synaptic densities (PSDs), which are located

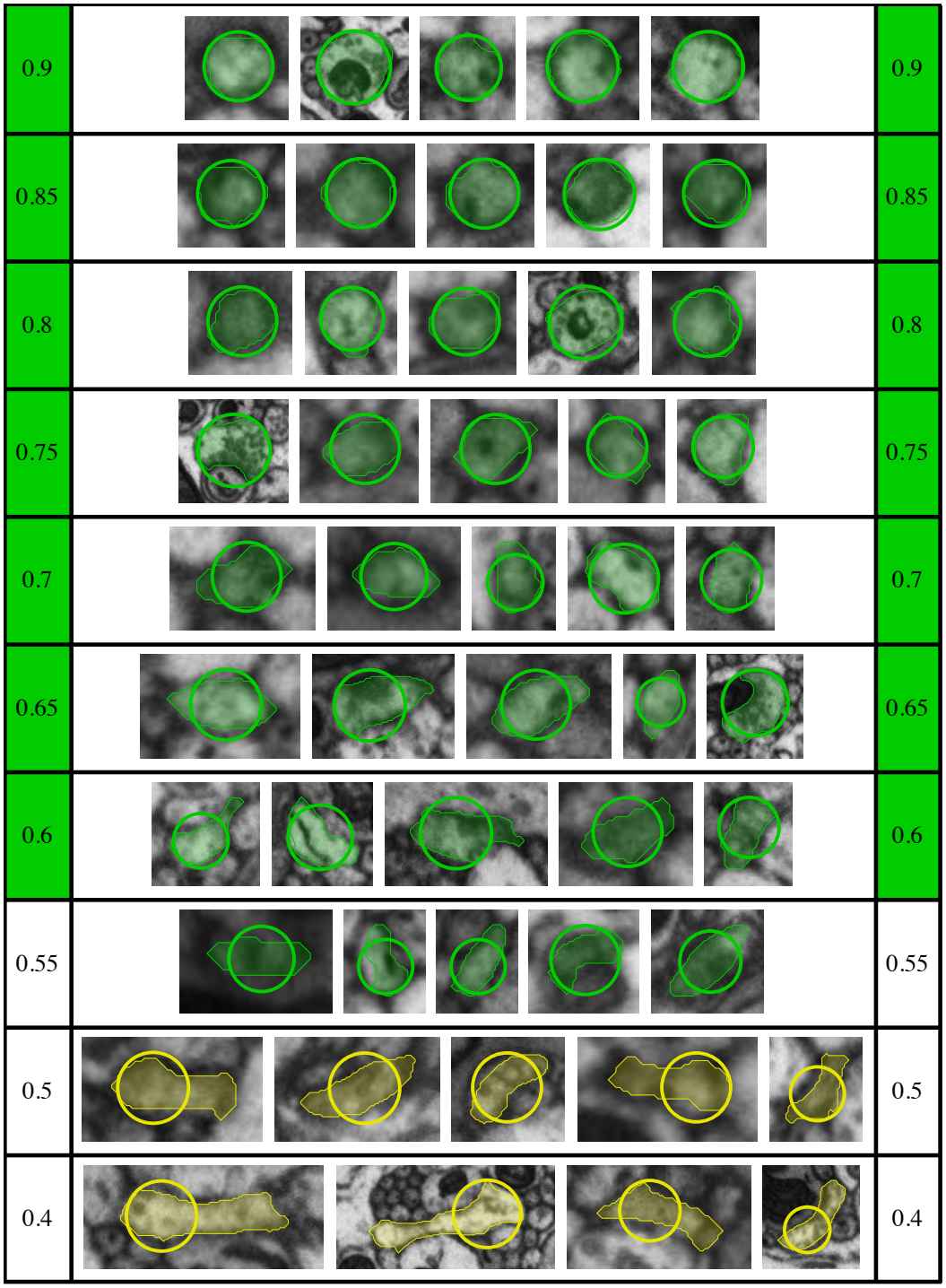


Figure 5.3: Example pairs of fibre polygons and corresponding maximum overlap circles. Each row is a random sample of 5 fibre-circle pairs whose overlap is within ± 0.005 of the value indicated in the sidebars for the row. Polygon and circle colouring indicates *medium* (yellow) and *high* (green) overlap. There are no fibre-circle pairs with *low* (red) overlap. Only four fibre-circle pairs with an overlap of 0.4 are shown due to space constraints. Green colouring of sidebars indicates the mean human judgment for the *acceptable fit* overlap range for polygon-circle pairs. Circles are fitted by selecting those having the highest *overlap* with the polygons. This appears to work well, with most *maximum overlap* circles judged an acceptable fit by neuroscientists. Note that polygons are dilated by two pixels to include membrane prior to circle fitting, but undilated polygons are shown in this figure. Image scale varies with fibre size.

at contact points between two neurons where synapses are present. We consider how these automatically labelled PSDs might be automatically attributed to the neurons forming the associated synapses. If we consider the ideal pair of scenarios where we have both the true fibre cross-section polygons and their best fitting circles, it is clear that synapse attribution will be unambiguous in the polygon case. Assuming synaptic PSDs are labelled with 100% accuracy, attributing them to a pair of neurons can be done simply by identifying the two adjacent polygons contacting each PSD. On the rare occasion that a PSD contacts polygons representing more than two neurons, a human can be asked to make the assignment. In the circle case, there may be many cases where labelled PSDs do not contact one or both of the circles they should be associated with. However, it is not clear how much more ambiguous this makes the attribution process. PSDs are fundamentally 2D structures and have a characteristic plane that is parallel to the membrane separating the two neurons forming the synapse. Therefore, attributing PSDs to the two closest circles along the axis perpendicular to the synaptic plane is likely to be unambiguous in many cases (see figure 5.4). Most synapses in our volume involve a cellular substructure called a *dendritic spine*. These are thin protrusions from the main trunk of a dendrite that can travel some distance to meet an axon and make a synapse. While many spines have cross-sections that are well represented by a circle (see figure 5.4), spine geometry can also be relatively complex and not well represented by a circle. If no circle representation of the post-synaptic spine is found, the "closest perpendicular circles" approach described above is likely to fail. Another case where synapse attribution is likely to fail is for synapses between axons and the trunks of dendrites (red line in figure 5.4), as dendrites are not well represented by circles in our chosen image plane. How successful the "closest perpendicular circles" approach is will crucially depend on the separation between nearby dendrites and the geometry of their spines, and further analysis will be necessary to determine this. In this work we restrict ourselves to the reconstruction of axonal fibres, and do not tackle the problem of synapse identification or attribution. However, any analysis of the connectivity in this region of the brain will eventually need to solve the problem of synapse identification, including the challenge of accurately reconstructing dendrites and their spines.

Branching: It may be possible to directly detect branch points from characteristic image features. In this case, it is possible that pre- and post-branch neurite segments might connect with detected branch points in both polygon and circle representations, making branch tracking relatively straightforward. Where branch points are not directly detected, one approach to accurately track a neurite as it splits to form two branches would be to associate a pair of post-branch cross-sections with a single pre-branch cross-section on the basis of their overlap. For

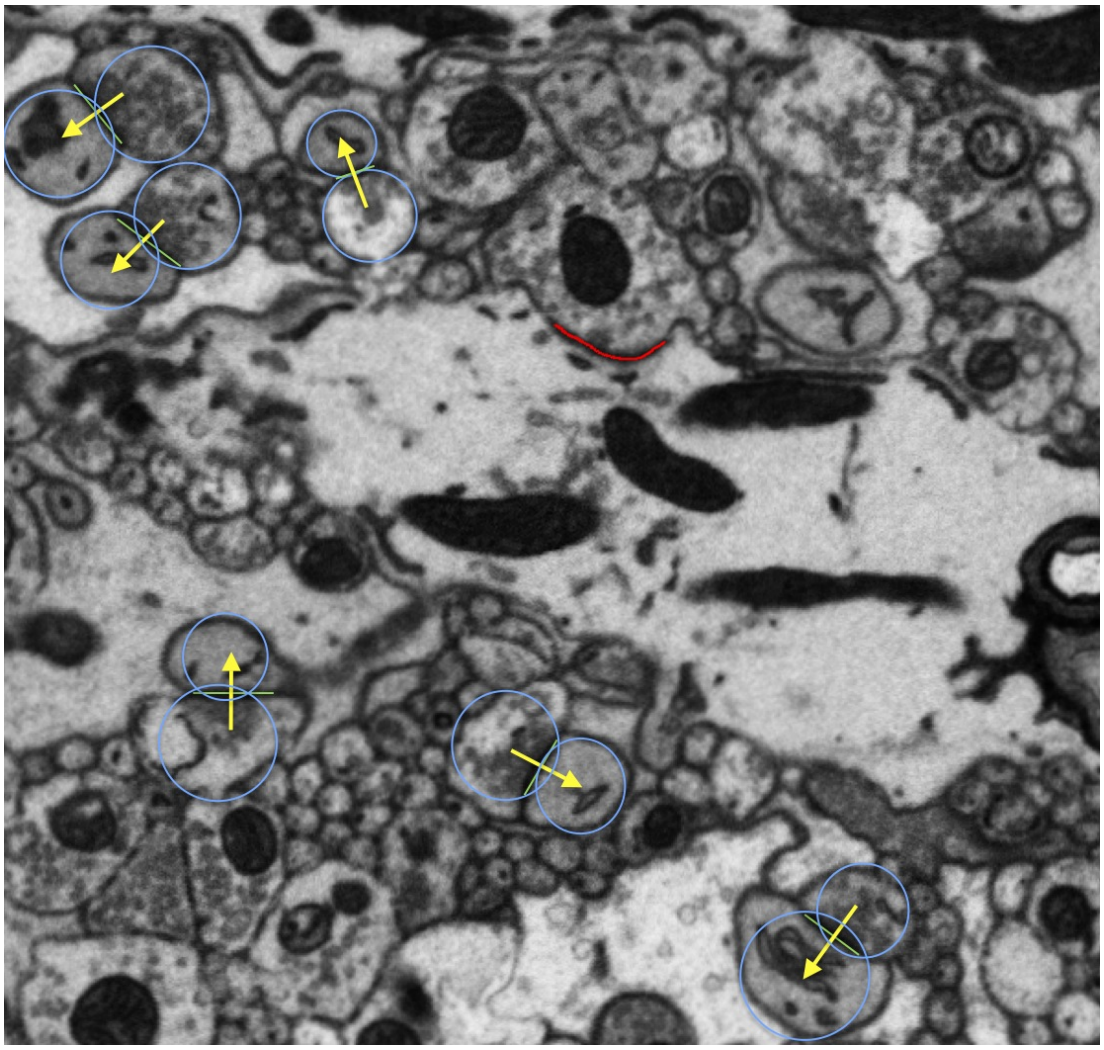


Figure 5.4: Illustration of why synapse attribution is likely to be unambiguous in most cases even when fibre cross-sections are represented as circles. Five synapses between parallel fibre boutons and dendritic spines are present in this image. The circular approximations for the pre- and post-synaptic cross-sections are shown in blue. The post-synaptic density (PSD) marking the synapse can be seen as a thick dark region along the membrane between the two cells. In all cases, the PSD is essentially a 2D structure, with two "thick" dimensions along the membrane surface and one "thin" dimension perpendicular to the membrane surface. Each PSD is marked with a thin green line, the length of which marks the extent of the in-plane "thick" dimension of the PSD. The second thick dimension of each PSD extends perpendicular to the image plane, while the final "thin" dimension is in-plane (perpendicular to and not much thicker than the green line). Yellow arrows have been drawn along the "thin" dimension of each PSD (perpendicular to the synaptic plane) and it can be seen that the nearest circles in all cases are cross-sections belonging to the true pre- and post-synaptic cells. Note that many spine cross-sections involved in synapses are well represented by circles in our chosen image plane. However, elsewhere in the volume there are spine cross-sections that are not well represented by circles. Synapse attribution is likely to fail in these cases, as there will be no post-synaptic circle to assign the synapse to. Another case where synapse attribution will fail is for synapses between an axon and the trunk of a dendrite. One such synapse appears in this image, and the red line near the centre of the image indicates the PSD for this synapse. In our chosen image plane the cross-section of the dendrite is not well represented by a circle and there would therefore be no post-synaptic circle to assign the synapse to. It may be that these dendrites and spines may be well represented by a set of circular cross-sections if the 3D image volume is "re-sliced" along a different axis. However, this is beyond the scope of this work.

a polygon representation, it is likely that both post-branch cross-sections will almost always overlap significantly with their pre-branch cross-section. To understand potential branching issues associated with the circle representation we consider the example polygon-circle pairs shown in 5.3. Looking at the high overlap pairs (green) it is likely that both post-branch circles will almost always overlap significantly with their pre-branch circle. However, looking at the medium overlap pairs (yellow), it is clear that the circles representing some elongated fibre cross-sections are often positioned to one side of their associated polygon. If branching was to occur at this point, it is unlikely that both post-branch circles would overlap with their pre-branch circle. It should be noted that medium overlap polygon-circle pairs represent only 1.2% of all fibre cross-sections in our data set. When coupled with the fact that not all of these elongated cross-sections will be at branch points, it seems reasonable to conclude that any branching issues associated with a circle representation will likely be limited to a small subset of branch points. It may also be possible to automatically identify these branch tracking failures for human review by identifying morphology that is not biologically plausible (e.g. a stretch of neurite that is isolated from a cell body). However, in this work we focus on the reconstruction of non-branching parallel fibres and therefore restrict ourselves to the problem of reconstructing neurite cross-sections and do not consider the issue of tracking branching neurites.

Neural network abstractions

Binary connectivity matrix: The simplest representation of a network of neurons is the information about which neurons are connected to each other. For a network of n neurons this comprises an $n \times n$ binary connectivity matrix C. An element c_{ij} is 1 if there is a synapse from neuron i to neuron j and 0 if there is not. As chemical synapses are unidirectional, elements c_{ij} and c_{ji} do not have to be identical. A second matrix could also be made for electrical synapses (gap junctions). These are bidirectional, so elements c_{ij} and c_{ji} would be identical. While any issues with the attribution of synapses to neuron pairs or the tracking of branches will obviously strongly affect the accuracy of the connectivity matrix, the reconstructed fibre cross-sections are not actually used to generate this representation and therefore circles are no less useful than polygons.

Weighted connectivity matrix: As with the binary connectivity matrix, a network of n neurons is represented by an $n \times n$ connectivity matrix C. However, each element c_{ij} now contains a real-valued representation of the absolute or relative strength of the connection from neuron i to neuron j. For a matrix representing connectivity via chemical synapses, these *connection weights* can also be negative, indicating inhibitory synapses. As with the binary matrix, c_{ij} is 0 if there is no connection from neuron i to neuron j. Helmstaedter et al. (2013) generate and

analyse such a weighted connectivity matrix, using the surface area of inter-neurite contacts as a proxy for connection strength. Such weighted connectivity matrices can also be used to model activity in networks of point neurons with either rate coded or spiking activity. Again, while any issues with the attribution of synapses to neuron pairs or branch tracking will obviously strongly affect the accuracy of the connectivity matrix, the reconstructed fibre cross-sections are not actually used to generate this representation and therefore circles are no less useful than polygons.

Wireframe models of neurons: In a wireframe model, each neuron is represented by a one-dimensional branching tree, with synapses between neurons located along these branches. With such models the morphology of neurons can be analysed, as can any patterns in the arrangement of synapses between neurons. A recent analysis of connectivity between two classes of retinal cells was able to identify selective connectivity by combining functional imaging and such wireframe models of pairs of neurons (Briggman, Helmstaedter, and Denk, 2011). A circle representation would have no impact on the topology of the reconstructed neuron trees. However, differences in the centres of mass estimated from polygon and circle based reconstructions could impact any analysis that relied on estimates of distances along the trees. To assess the size of any impact of a circle representation on tree length estimates, we calculated the relative difference between the total lengths obtained using ground truth polygons and maximum overlap circles for 439 parallel fibre segments traversing the $\sim$5 μm z-extent of our volume. The relative length difference is defined as $\frac{|length_{circle} - length_{polygon}|}{length_{polygon}}$ and the maximum difference in fibre length between the two representations is less than 1%, with the median difference being less than 0.1%. We would therefore suggest that circles are no less useful than polygons for generating wireframe models of neurons.

Electrical models of neurons: Single and multi-compartment models can be used to model the passive and dynamic electrical properties of neurons. Such models can capture the influence of synapse and branch geometry on the summation of membrane potentials and the generation and propagation of dendritic spikes. One notable project using such models is *The Blue Brain Project* (Markram, 2006), which is modelling thousands of neurons to simulate rat whisker sensing circuits. Most simulations using electrical models of neurons approximate sections of neurites as either cylinders or cone segments with circular cross-sections (e.g. NEURON: Carnevale and Hines, 2006). We would therefore suggest that circles might be no less useful than polygons for most electrical models of neurons.

Low-level models of sub-cellular neural processes: Some very detailed models of neurons incorporate low-level sub-cellular processes such as neurotransmitter diffusion (e.g. M-Cell:

Stiles and Bartol, 2001). For such models a circular representation may be inadequate if precise surface meshes are required for accurate modelling.

5.3 Finding representative circles from overlap

5.3.1 Ground truth overlap as a "fibreness" score

Having established that circles are an appropriate representation for fibre cross-sections, we must now determine how to find the most appropriate set of circles to represent a given set of fibre cross-sections. To do this we introduce the concept of the *ground truth overlap* between a given circle and the set of ground truth polygons representing a collection of fibre cross-sections. This is calculated by pairing the circle with all overlapping fibre polygons. The pairing with the largest overlap is selected and this overlap is defined as the *ground truth overlap* for the circle. This *ground truth overlap* can be interpreted as a "fibreness" score indicating how well a given circle represents any of the ground truth polygons. By repeating this process for circles with a range of radii at every pixel location within the ground truth, we generate the *ground truth overlap volume* for the set of ground truth polygons. This is a 3D *x,y,r* volume indicating how well each candidate circle is supported by the ground truth polygons. Figure 5.5 shows an example ground truth overlap volume as a 3D rendering. In practice, it is not necessary to consider every possible radii for candidate circles and we have found a non-uniform sampling of 20 radii between 6 and 78 pixels to be sufficient for accurate circle finding.

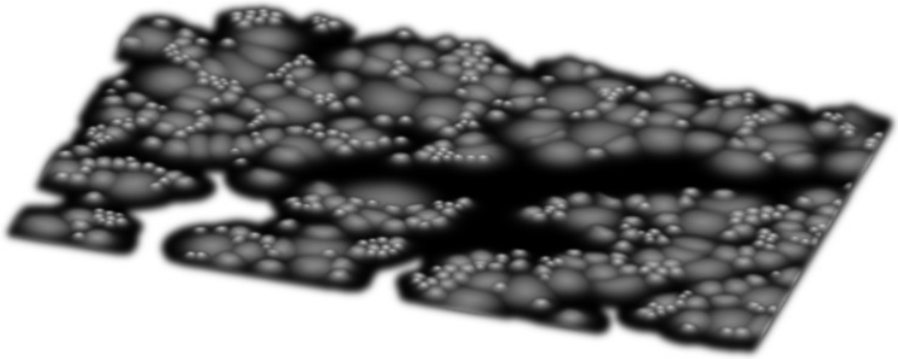


Figure 5.5: Ground truth overlap volume for the ground truth fibre cross-section polygons associated with a $1,274 \times 852$ electron microscope image of the molecular layer of the mouse cerebellum (data set 1171). The z-axis is non-linear, with each z-layer representing one of 20 non-uniformly sampled radii between 6 and 78 pixels (radii: 6, 7, 8, 9, 10, 11, 12, 14, 16, 18, 21, 24, 28, 32, 37, 43, 50, 58, 67, 78). Each point plots the *ground truth overlap* with the ground truth polygons for a circle with the corresponding position and radius.

5.3.2 Finding circles from ground truth overlap

Having generated a *ground truth overlap volume* for a given set of ground truth polygons we must now use it to generate a set of circles representing this ground truth. To do this, we greedily select the circles corresponding to the points in the volume with the largest *ground truth overlaps*. Given that the ground truth polygons are non-overlapping, we also apply a soft constraint to ensure our found circles do not excessively overlap (once we have selected a circle, we exclude any remaining candidate circles where the centre of the smaller circle is within half its radius of the edge of the larger circle). Having excluded these circles from consideration we then select the remaining candidate circle with the largest overlap. We repeat this process until we reach a predetermined minimum overlap threshold. This process is outlined in algorithm 1.

Algorithm 1: Finding circles from ground truth overlap

```
Data: Ground truth overlap at all positions for a range of circle radii; stopping overlap
      threshold.
Result: Predicted circles representing fibre cross-sections in ground truth.
while largest ground truth overlap > stopping threshold do
    select circle {x_f, y_f, r_f} with largest ground truth overlap;
    select circle as fibre and set ground truth overlap at {x_f, y_f, r_f} to zero;
    for all pairings of current circle and remaining candidate circles do
        if inter-circle distance < (larger radius + 0.5 × smaller radius) then
            set ground truth overlap for paired candidate circle {x_p, y_p, r_p} to zero;
        end
    end
end
```

We assess the quality of the set of found circles using the *overlap f-measure* introduced in section 3.5.3. To calculate the *overlap f-measure* we first greedily pair our found circles with the ground truth fibre polygons in order of decreasing *overlap* between pairs. We then calculate the overall *matched overlap* between the set of circles and the set of ground truth polygons by summing these pair-wise overlaps across all matched circle-polygon pairs. *Precision* and *recall* are then calculated by dividing the *matched overlap* by the numbers of found circles and ground truth polygons respectively. Finally, the *overlap f-measure* is calculated by taking the *harmonic mean* of *precision* and *recall*. This process is outlined in algorithm 2.

5.3.3 Predicting overlap is sufficient for finding good quality circles

Having established a method for finding circles from *ground truth overlap volumes*, we must determine whether these found circles are a suitable representation of the underlying ground truth fibre polygons. To do this we compare the distribution of polygon-circle overlaps for the circles we find with our algorithm to the distribution of overlaps for the circles derived directly

Algorithm 2: Calculating the overlap-based f-measure between sets of found circles and ground truth polygons.

Data: Set of found circles; set of ground truth fibre polygons.
Result: Overlap-based f-measure similarity measure.
for *all pairings of found circles and ground truth polygons* **do**
  calculate *overlap* between each polygon-circle pair;
end
while *pair with non-zero overlap remains* **do**
  select pairing with highest overlap;
  set overlap for all other pairs involving selected circle and polygon to zero;
end
calculate *matched overlap* by summing overlap for all selected polygon-circle pairs;
calculate *precision* by dividing *matched overlap* by number of found circles;
calculate *recall* by dividing *matched overlap* by number of ground truth polygons;
calculate *overlap f-measure* by taking the *harmonic mean* of *precision* and *recall*;

from the ground truth fibre polygons. Figure 5.6 shows these two distributions for comparison.

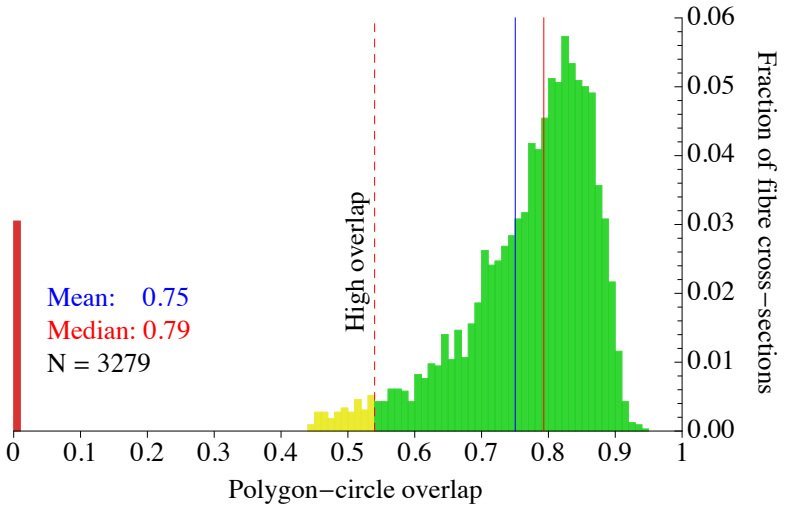
It can be seen that the two distributions are very similar and all circles found from the ground truth *ground truth overlap volume* were matched to true fibre polygons with medium or high overlap (94.3% high overlap vs. 98.8% for circles derived directly from the ground truth polygons). While there are 100 ground truth polygons for which no corresponding circle was found, these represent only 3.0% of true fibre polygons. Therefore, we would suggest that accurately predicting the *ground truth overlap volume* associated with a set of fibre cross-sections is sufficient to generate a suitable set of circles representing these cross-sections.

As a sanity check we also show the actual found circles from one data set (1171), along with the corresponding ground truth polygons (figure 5.7). The vast majority of found circles have high overlap with the ground truth polygons and, while a few polygons have no corresponding found circle, no erroneous circles are found.

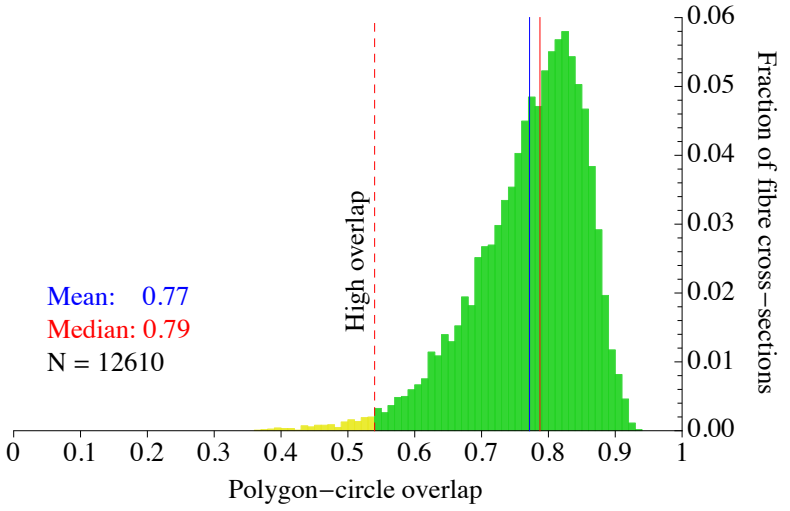
5.4 Predicting overlap from image features

5.4.1 Basic Image Features (BIFs)

In this work we will use *Basic Image Features* (BIFs: Crosier and Griffin, 2010) to quantify the evidence for fibre cross-sections encoded in the image. BIFs are a classification of local image structure into seven different classes on the basis of approximate local symmetry. The seven classes are *flat*, *gradient*, *dark blob*, *light blob*, *dark line*, *light line* and *saddle*. The BIF scheme can also incorporate feature orientation, and these augmented features are known as *oriented BIFs* (oBIFs; Lillholm and Griffin, 2008). Flat and blob oBIFs have no associated orientation, gradient oBIFs have one of eight unidirectional orientations and line and saddle oBIFs have one of four bidirectional orientations, giving 23 oBIF sub-classes. The key properties of the BIF

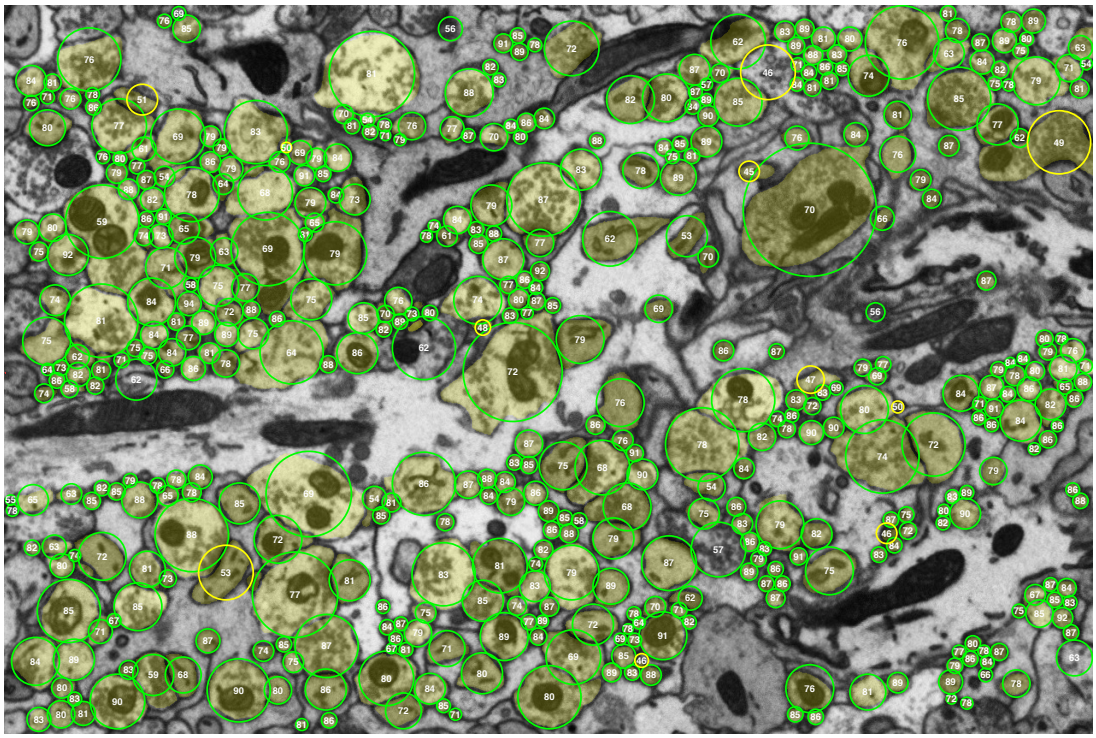


(a) Overlap distribution for circles found from the ground truth *maximal overlap* volume

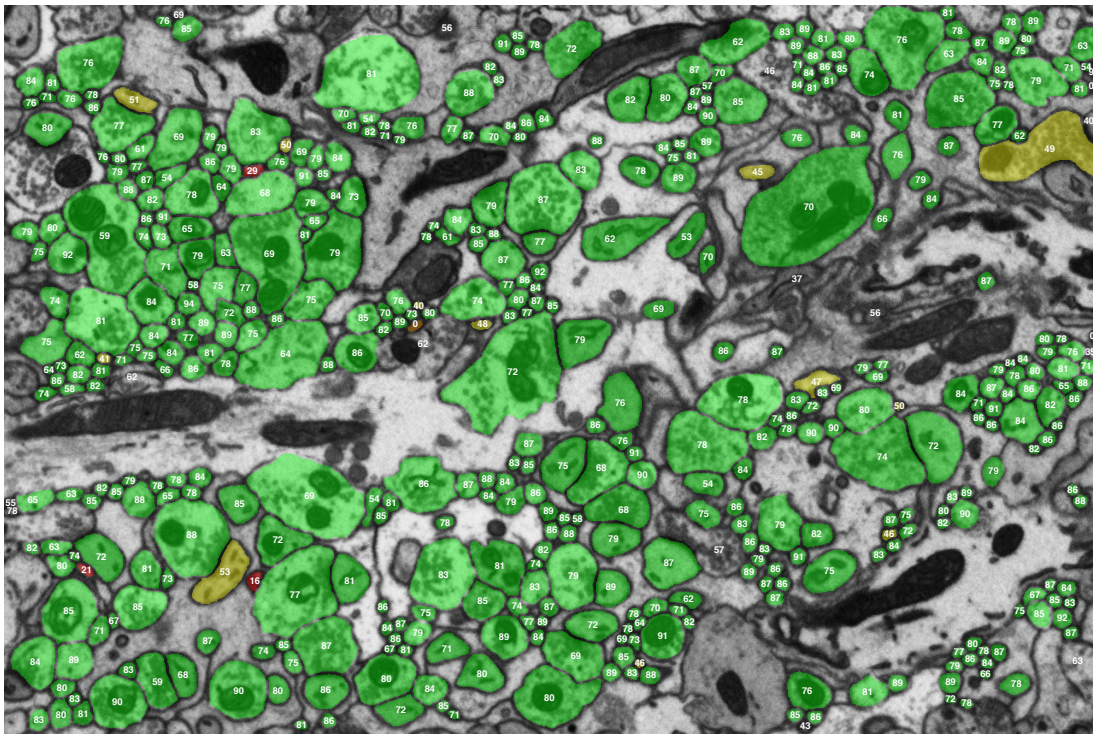


(b) Overlap distribution for circles derived directly from the ground truth polygons

Figure 5.6: Comparing overlap distributions for **(a)** circles found from the ground truth *ground truth overlap volume* and **(b)** circles derived directly from the ground truth fibre polygons (data as in figure 5.2). Bar colours indicate *low* (red), *medium* (yellow) and *high* (green) overlap. Mean and median overlaps are indicated with blue and red solid vertical lines. The two distributions are very similar, although the *ground truth overlap* distribution has more mass in the lower overlap range, including 100 true fibre polygons for which no overlapping circle was found (red bar at far left). Despite this, 94.3% of fibre polygons have high overlap with a circle found from the ground truth *ground truth overlap volume* (c.f. 98.8% for circles derived from ground truth polygons) and only 3.0% have low overlap. None of the circles found from ground truth *ground truth overlap volume* had low overlap. To generate the data for (b) the *stopping overlap threshold* for algorithm 1 was set from volume 1341 by maximising the *f-measure* defined in algorithm 2. This threshold was then used to find circles for data sets 1001, 1171, 1511, 2001, 2171, 2341 and 2511. The found circles and ground truth polygons were then pooled across data sets to generate the distribution in (b) and corresponding *precision* (0.77), *recall* (0.75) and *f-measure* (0.76) statistics.



(a) Fibre circles found using the ground truth *ground truth overlap* volume



(b) Fibre polygons from manually labelled ground truth

Figure 5.7: **(a)** Fibre circles found using the ground truth overlap volume, along with the underlying ground truth polygons for comparison. Circle colour coding indicates *high* (green), *medium* (yellow) and *low* (red) overlap with the ground truth polygons. Actual overlap percentages are indicated by white text. **(b)** The corresponding manually labelled ground truth polygons colour coded by overlap using the same green/yellow/red scheme to indicate high/medium/low overlap. Again, actual overlap percentages are indicated by white text.

classes are detailed in table 5.1.

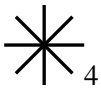
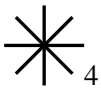
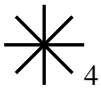
BIF class	Orientations	Optimal stimulus	Key	Response magnitude (R)
Flat	None			γL_{00}
Gradient	 8			$\sigma\sqrt{L_{10}^2 + L_{01}^2}$
Dark blob	None			$\frac{1}{2}\sigma^2(\lambda_{H1} + \lambda_{H2}) = \frac{1}{2}\sigma^2(L_{20} + L_{02})$
Light blob	None			$-\frac{1}{2}\sigma^2(\lambda_{H1}+\lambda_{H2}) = -\frac{1}{2}\sigma^2(L_{20}+L_{02})$
Dark line	 4			$\frac{1}{\sqrt{2}}\sigma^2\lambda_{H1}$
Light line	 4			$-\frac{1}{\sqrt{2}}\sigma^2\lambda_{H2}$
Saddle	 4			$\frac{1}{2}\sigma^2(\lambda_{H1} - \lambda_{H2}) = \frac{1}{2}\sigma^2\sqrt{(L_{20} + L_{02})^2 + 4L_{11}^2}$
$\lambda_{H1} = \frac{1}{2}(L_{20} + L_{02} + \sqrt{(L_{20} + L_{02})^2 + 4L_{11}^2})$ $\lambda_{H2} = \frac{1}{2}(L_{20} + L_{02} - \sqrt{(L_{20} + L_{02})^2 + 4L_{11}^2})$				

Table 5.1: BIF classes: Gradient oBIFs have eight unidirectional orientations. Line and saddle oBIFs have four bidirectional orientations. L_{nm} is the output after convolving the image with a Derivative of Gaussian filter of order n in the x-direction and order m in the y-direction. σ is the standard deviation of the Derivative of Gaussian filters. γ is a multiplier that determines when the *flat* BIF response will dominate over those of the other BIF classes. λ_{H1} and λ_{H2} are the largest and smallest eigenvalues of the Hessian matrix of the smoothed image: $H = \left[\begin{smallmatrix} L_{20} & L_{11} \\ L_{11} & L_{02} \end{smallmatrix}\right]$.

BIF classes are determined by the output of a family of *Derivative of Gaussian* filters (table 5.2). At every pixel the response for each BIF class is calculated as in table 5.1. The pixel is then assigned to the BIF class with the largest response. The use of *Derivative of Gaussian* filters as a basis for image representation was first proposed by Koenderink (1984), and such filters have been used to model both retinal and cortical receptive fields in the mammalian visual system (Young, 1987; Young, Lesperance, and Meyer, 2001). These filters are also *steerable*, meaning that the response of a filter at any orientation can be calculated as a linear combination of the partial derivative filters (Freeman and Adelson, 1991). The family of partial derivative filters up

to order *n* is called the *n-jet* and for BIFs we use the *2-jet* (zeroth, first and second orders). The filters have three parameters: the scale of the gaussian (σ: standard deviation; common across the family of filters) and the partial derivative orders in x and y (n and m; different for each filter). σ determines how much the image is blurred prior to taking its derivatives. The larger σ, the larger the features in the image must be to produce a strong response for the corresponding BIF classes. The BIF scheme has one additional parameter γ, which is a multiplier for the *flat* BIF response that determines when it dominates over the responses of the other BIF classes.

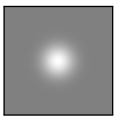
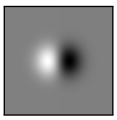
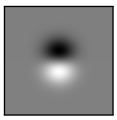
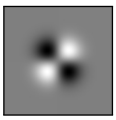
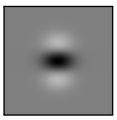
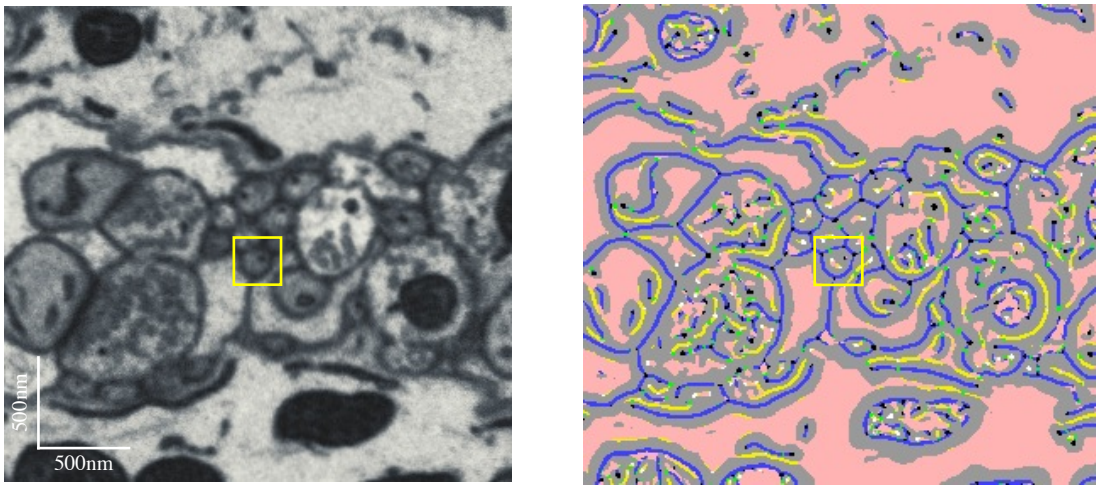
Second-order family of Derivative of Gaussian filters					
					
F_{00}	F_{10}	F_{01}	F_{20}	F_{11}	F_{02}

Table 5.2: Second-order family of Derivative of Gaussian filters. F_{nm} is a filter of order n in the x-direction and order m in the y-direction. The scale of the filter family is controlled by the standard deviation of the Gaussian (σ).

The BIF classes share much in common with other sets of image features used in the field of computer vision. The response of the *gradient* BIF is simply the magnitude of the *gradient* of the blurred image, while the responses of the second-order *blob*, *line* and *saddle* BIFs are linear combinations of the eigenvalues of the *Hessian matrix* of the blurred image. The *Hessian matrix* of an image is the matrix of its partial second-order derivatives (in our case $H = \left[\begin{smallmatrix} L_{20} & L_{11} \\ L_{11} & L_{02} \end{smallmatrix}\right]$, where L_{nm} is the response of filter F_{nm}). Both the gradient magnitude and the Hessian eigenvalues are also used by *ilastik*, the pixel-based classifier we benchmark our model-based approach against (Sommer et al., 2011). Histograms of gradient oBIFs are also widely used by others under the moniker *histogram of gradients* (HoG). While the component BIF classes may not be unique to the field, differences with other feature schemes include the early hard classification of pixels and the use of histograms of BIFs rather than vectors of feature responses. BIF class response magnitudes are also normalised to have approximately the same response to their optimal stimuli. However, BIFs were primarily selected because they are a principled family of features that correspond to elements of interest for the task of identifying fibre cross-sections in EM images (see section 5.4.2). Nonetheless, we do not claim that the BIF scheme is uniquely suited for this task.

5.4.2 Annular BIF histograms as "fibreness" feature vectors

Figure 5.8 shows the EM image for a region of the molecular layer of mouse cerebellar cortex and the corresponding BIF class for each pixel at the optimal σ (1.75) and γ (0.085) for fibre detection in our circle-based approach (section 6.3.1). The external membrane of fibre cross-sections is dominated by *dark line* (blue) and *gradient* (grey) BIFs. However, these BIFs are also present where there is interior membrane. It can be seen in figure 5.3 that, for fibres with high overlap circle representations, the external membrane lies close to the perimeter of the associated circle. We therefore opted to use an *annulus* as our circle model in order to capture the features associated with external fibre membrane while excluding those associated with interior membrane.



(a) EM image of a region of mouse cerebellum

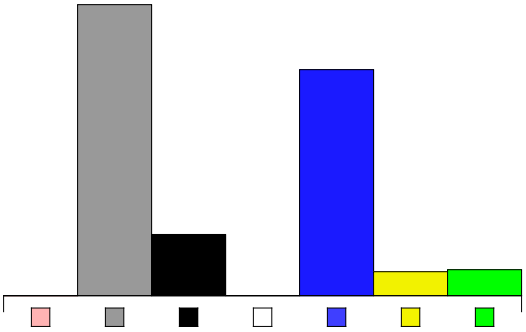
(b) Corresponding BIF classes

Figure 5.8: **(a)** EM image and **(b)** corresponding BIFs for a region of mouse cerebellum. Colour key for BIFs is as specified in table 5.1. The yellow box indicates the fibre neighbourhood shown in figures 5.9 and 5.10. The external membrane of fibre cross-sections is dominated by *dark line* (blue) and *gradient* (grey) BIFs. However, these BIFs are also present where there is interior membrane.

Figure 5.9 shows how we generate the histogram of BIFs for the annulus associated with a fibre. The selection of optimal BIF and histogram parameters is discussed in section 6.3. We set the inner and outer radii of the annulus to 0.6 and 1.3 times the radius of the circle that represents the fibre (section 6.3.1). We divide the annulus into 8 segments and generate the histogram of BIFs for each segment. We then take the square root of these segment histograms (section 6.3.2). This is a common approach in computer vision and makes the elements of the histogram have equal precision. Finally, we take the mean of these square-rooted histograms to obtain the overall histogram for the fibre. By averaging histograms over segments we hope to

enforce a requirement for consistent membrane evidence along the perimeter of each putative fibre.

(a) Fibre image (b) Fibre BIFs (c) Fibre annulus segments

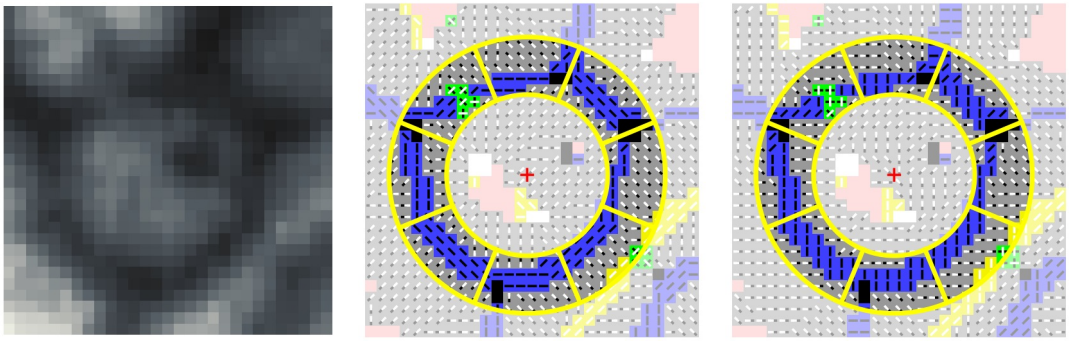


(d) Fibre annular BIF histogram

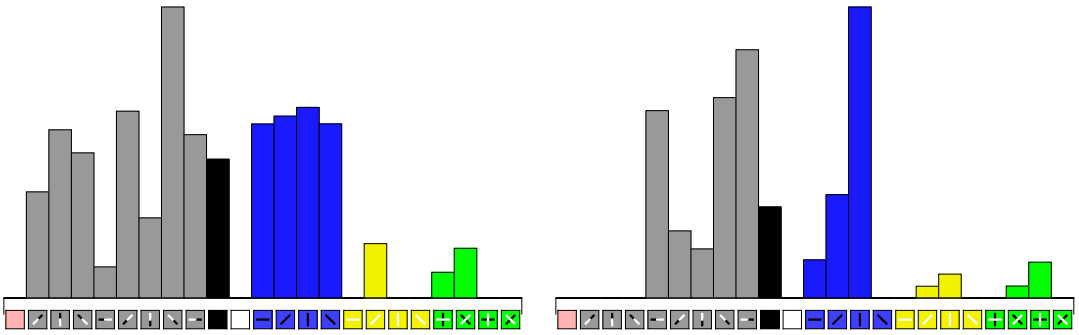
Figure 5.9: Annular BIF histogram. **(a)-(c)** Fibre EM image and corresponding BIFs and annular segments. **(d)** The corresponding BIF histogram for the fibre generated by taking the mean of the square-rooted BIF histograms for each of the 8 annular segments.

We can incorporate further expectations derived from our circle representation into our BIF histogram scheme. As most fibre cross-sections are approximately circular, we expect that the *orientation* of the informative oBIFs will be consistent in a polar reference frame. We expect that *gradient* (grey) oBIFs will be oriented approximately parallel or anti-parallel to the radial vector from the centre of the annulus to the BIF pixel location and that *dark line* (blue) oBIFs will be oriented approximately perpendicular to this vector (figure 5.10b). We can encode this expectation by extending the BIF scheme to normalise the orientation of oBIFs within an annulus to be relative to the radial vector from the annulus centre to the BIF pixel location. We call this extended scheme *radially normalised oBIFs* (rBIFs). Smith, Carleton, and Lepetit (2009) use a similar orientation normalisation approach for the image gradient when calculating their *ray features*. Figure 5.10 shows how we generate the rBIF histogram for a fibre and compares oBIF and rBIF histograms. For efficiency reasons, we quantise the

oBIF orientations prior to radially normalising them. In order to mitigate the quantisation noise introduced by this process we perform a soft quantisation, assigning fractions of a pixel to each quantised orientation depending on the angle between the unquantised BIF orientation and each quantised orientation. The fraction of a pixel assigned to each quantised orientation is calculated using the formula $\frac{1}{\sqrt{2\pi}\sigma} \exp -\frac{(\theta-\theta_Q)^2}{2\sigma^2}$, where θ is the unquantised orientation and θ_Q is the quantised orientation. This is a 1D Gaussian with a mean of θ_Q and a standard deviation of σ. In this work we set σ to 0.65.



(a) Fibre image (b) Fibre annulus and oBIFs (c) Fibre annulus and rBIFs



(d) Fibre oBIF annular histogram (e) Fibre annular rBIF histogram

Figure 5.10: Radially normalised BIFs (rBIFs): **(a)-(c)** Fibre EM image and corresponding oBIFs and rBIFs with segmented annuli. rBIFs are generated from oBIFs by normalising the BIF orientation to be relative to the vector from the centre of the annulus to the pixel location of the BIF. **(d)-(e)** The corresponding oBIF and rBIF histograms for the fibre generated by taking the mean of the square-rooted BIF histograms for each of the 8 annular segments. rBIF histograms for fibres are "spikier" than oBIF histograms, reflecting the increased consistency of BIF orientation after radial normalisation.

Chapter 6

Reconstructing fibre cross-sections

In chapter 5 we introduced a circle representation of fibre cross-sections and demonstrated that predicting the *ground truth overlap* of candidate circles was sufficient to generate a high quality reconstruction. In this chapter we describe how we learn to predict this overlap from electron microscope images, using annular histograms of Basic Image Features (BIFs) to assess the image evidence for each circle. We then describe our process for finding a representative set of circles from this predicted overlap and discuss the selection of training data, algorithm parameters and prediction method. Finally, we evaluate the performance of our algorithm and benchmark it against *ilastik*, a state of the art pixel-based classifier.

6.1 Algorithm overview

Figure 6.1 presents an overview of our process for learning to predict a set of circles representing the fibre cross-sections present in a 2D electron microscope (EM) image. It illustrates three main components: the "fibreness" feature vector, the "fibreness" score and the fibre finding process. We describe the algorithm in more detail below.

6.1.1 Learning a mapping from EM image to ground truth overlap

In section 5.3 we defined *ground truth overlap* as the maximal overlap between a candidate circle and the set of fibre cross-sections in the ground truth. We demonstrated that predicting the *ground truth overlap* for all possible candidate circles is sufficient to generate a high quality reconstruction by greedily selecting the set of non-overlapping candidate circles with the highest overlaps. In section 5.4 we introduced BIFs as a family of image features and annular BIF histograms as a "fibreness" feature vector encoding the underlying image evidence for fibre circles. The first stage of our algorithm is to learn a regression function encoding a mapping between the annular BIF histogram associated with each circle and its corresponding *ground truth overlap*. This process is illustrated in figure 6.1 (a-g) and corresponding pseudocode is provided in algorithm 3. To learn this mapping we select training data, generate annular BIF histograms and

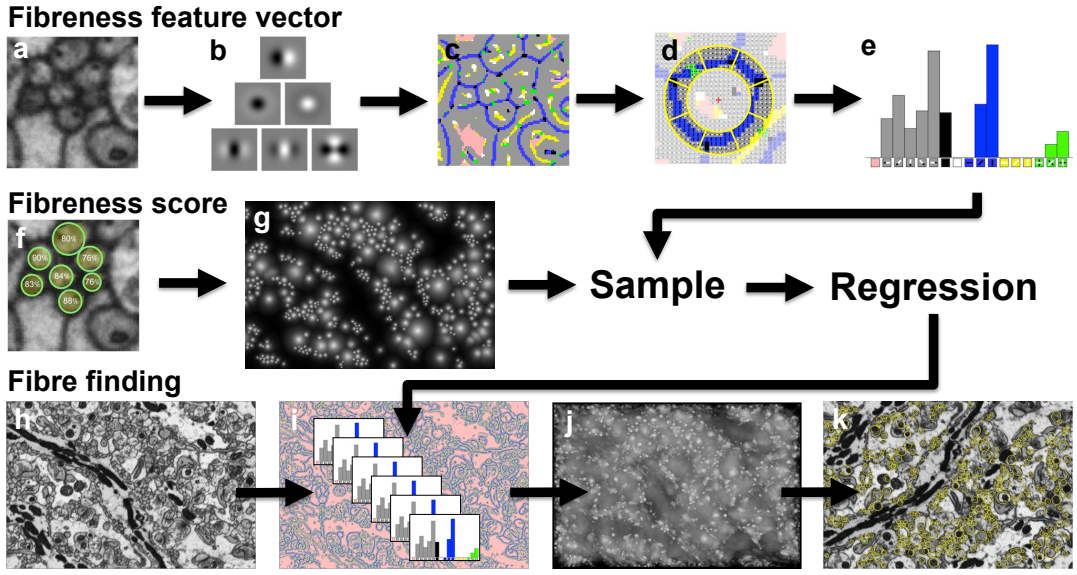


Figure 6.1: Overview of 2D fibre finding algorithm. **Top row:** An EM image (a) is processed by a bank of Derivative-of-Gaussian filters (b) to generate an oBIF map (c). For each possible circle, the subset of oBIFs falling within an associated annulus is selected. oBIF orientations are normalised to be relative to the radial vector from the centre of the circle to each pixel, generating rBIFs (d). The mean histogram of rBIFs across 8 annular segments is used as a feature vector to assess the evidence provided by the image for each circle (e). **Middle row:** The overlap with the ground truth is calculated for each possible circle (f). This is done for a range of radii at every pixel, generating a 3D *ground truth overlap volume* (g: maximum overlap over all radii). A sample of circles is selected and a regression function is learned to predict circle ground truth overlap from rBIF histograms. **Bottom row:** To find fibres, an EM image (h) is processed to generate annular rBIF histograms for all possible circles (i: image BIFs in background). These are processed using the learned regression to predict the ground truth overlap for all possible circles (j: maximum predicted overlap over all radii). This predicted overlap is used to place circles in a greedy, mutually exclusive manner (k).

then learn a regression function to predict *ground truth overlap* from these histograms. These steps are described below.

Select training data

We select a 1,274×852 pixel EM image and its associated ground truth fibre polygons as a training data set. We consider 20 possible radii for fibre circles at every pixel and generate the associated *ground truth overlap volume* for the ground truth fibre polygons (f-g). From this set of ∼19 million circles we select ∼65,000 circles as our training set. We found that a targeted sampling was more effective than a random sampling and we select our training circles using a two-stage process. We summarise this process below and discuss it further in section 6.2.

1. We select a maximal sample of circles such that the distribution of radii for our sample circles matches the distribution of radii for the circles that best represent true fibre cross-sections.

Algorithm 3: Learning a mapping from EM image data to circle ground truth overlap

```
Data: 2D electron microscope image; set of sample circles {x,y,r}.
Result: Function mapping annular BIF histograms to ground truth overlap.
generate BIF map for image using algorithm 9 (appendix C);
for all sample circles do
    determine ground truth overlap for circle;
    divide annulus associated with current circle into 8 equal segments;
    for each annulus segment do
        select BIFs within each annular segment;
        if BIF type is rBIF then
            set BIF orientations to be relative to radial vector from annulus centre to
            pixel;
        end
        if BIF type is oBIF or rBIF then
            quantise BIF orientations using algorithm 10 (appendix C);
            bin BIFs by class and orientation to generate segment BIF histogram;
        else
            bin BIFs by class to generate segment BIF histogram;
        end
        if normalisation type is square-root then
            re-normalise histogram by taking the square-root of each bin;
        end
    end
    generate the mean histogram across segments;
end
train regression function with mean histograms as input and ground truth overlaps as
target;
```

2. From this we select a sample of approximately 65,000 circles such that the distribution of *ground truth overlap* for our sample circles is approximately uniform.

Generate annular BIF histograms

We generate a BIF feature map for our training image (a-c) by convolving the image with a family of Derivative of Gaussian filters (table 5.2) and determining the BIF class with the maximal response at each pixel (table 5.1). For each of our training circles we then generate an associated annular BIF histogram (d-e) by considering the BIFs lying within each segment of an 8-segment annulus co-centred with the circle. We generate a histogram of BIFs for each segment and then take the average histogram across segments as the "fibreness" feature vector associated with the circle. For BIFs each histogram has one bin for each of the 7 top-level BIF classes. For oBIFs, each top-level BIF class can have one of up to 8 quantised orientations (see table 5.1), giving 23 histogram bins. For rBIFs these orientations are normalised to be relative to the radial vector from the centre of the annulus to each BIF pixel, and there are also 23 histogram bins. The selection of optimal BIF, annulus and histogram parameters is discussed

further in section 6.3.

Learn a regression function

Finally we learn a regression function to predict each circle's *ground truth overlap* from its annular BIF histogram. We have found learning a *logistic* regression function to be most appropriate for this task. This choice is discussed further in section 6.4.

6.1.2 Finding fibre circles

The process of finding fibre circles for a previously unseen EM image is illustrated in figure 6.1 (h-k). First we generate BIF histograms for 20 different circles at every pixel (h-i). We then use our learned regression function to predict the *ground truth overlap* for each of these circles (i-j). Finally, we use this predicted *ground truth overlap volume* to place a set of circles that hopefully represent the underlying true fibres (j-k). The generation of a predicted *ground truth overlap volume* from EM image data and the placement of fibre circles are described in algorithms 4 and 5 respectively. Algorithm 5 is identical to algorithm 1 except for the addition of a *luminance threshold* for mitochondria exclusion. This threshold is fixed globally for all data sets and requires the mean image luminance within any candidate fibre circle to be greater than 24% of the maximum luminance. The inclusion of this threshold greatly reduces the number of circles placed within mitochondria but does not entirely eliminate them. This luminance threshold was chosen to maximise the f-measure on the test data set used for selecting the BIF histogram parameters and regression method. This data set was not used for the final evaluation and benchmarking of 2D reconstruction performance.

6.2 Training data selection

6.2.1 Range of circle radii

The approach of our algorithm is to consider a range of circles that might correspond to fibre cross-sections in an EM image. However, we do not need to consider every possible circle radius. The range of radii for the circles best representing fibres in our ground truth data is 6-78 pixels, so we only need to consider circles with radii within this range. In practice, it is also not necessary to consider every possible radii within this range. We have found that generating a *ground truth overlap volume* for a non-uniform sampling of 20 radii between 6 and 78 pixels to be sufficient for accurate circle finding (section 5.3; radii = 6, 7, 8, 9, 10, 11, 12, 14, 16, 18, 21, 24, 28, 32, 37, 43, 50, 58, 67, 78 pixels). Our set of potential training circles therefore comprises 20 possible circles at every pixel in our training image. For a training image of 1,274x852 pixels, this gives $\sim$ 19 million circles that are fully supported within the image.

Algorithm 4: Predicting ground truth overlap from EM image data

Data: 2D electron microscope image; list of circle radii.
Result: Predicted ground truth overlap for all radii at all pixels.
generate BIF map for image using algorithm 9 (appendix C);
for *all circle radii at all pixels* **do**
  Divide annulus associated with current circle into 8 equal segments;
  for *each annulus segment* **do**
    select BIFs for pixels within segment;
    if *BIF type is rBIF* **then**
      set BIF orientations to be relative to radial vector from annulus centre to pixel;
    end
    if *BIF type is oBIF or rBIF* **then**
      quantise BIF orientations using algorithm 10 (appendix C);
      bin BIFs by class and orientation to generate segment BIF histogram;
    else
      bin BIFs by class to generate segment BIF histogram;
    end
    if *normalisation type is square-root* **then**
      re-normalise histogram by taking the square-root of each bin;
    end
  end
  generate the mean histogram across segments;
  predict ground truth overlap from mean histogram using trained regression function;
end

Algorithm 5: Finding circles from ground truth overlap prediction

Data: Predicted overlap at all pixels for a range of circle radii; stopping overlap threshold; EM image; mitochondria exclusion luminance threshold.
Result: Predicted circles representing fibre-cross-sections present in EM image.
while *maximum predicted overlap* $>$ *stopping threshold* **do**
  select circle $\{x_c, y_c, r_c\}$ with largest predicted overlap;
  calculate mean image luminance within circle;
  if *mean luminance* $<$ *mitochondria threshold* **then**
    reject circle as mitochondrion and set predicted overlap at $\{x_c, y_c, r_c\}$ to zero;
  else
    select circle as fibre and set predicted overlap at $\{x_c, y_c, r_c\}$ to zero;
    for *all pairings of current circle and remaining candidate circles* **do**
      if *inter-circle distance* $< (larger\ radius + 0.5 \times smaller\ radius)$ **then**
        set predicted overlap for paired candidate circle $\{x_p, y_p, r_p\}$ to zero;
      end
    end
  end
end

6.2.2 Picking training circles

We have found that taking a random sample of the ∼ 19 million possible training circles does not give us the best fibre finding performance. Instead, biasing our sample in two key ways results in improved performance.

Sample bias 1: Mirroring the distribution of radii for ground truth fibre circles

The distribution of radii for ground truth fibre circles is highly non-uniform (figure 6.2). The minimum radius is 6 pixels and 78% of fibre circles have a radius of 16 pixels or less, with 89% of fibre circles having a radius of 24 pixels or less.

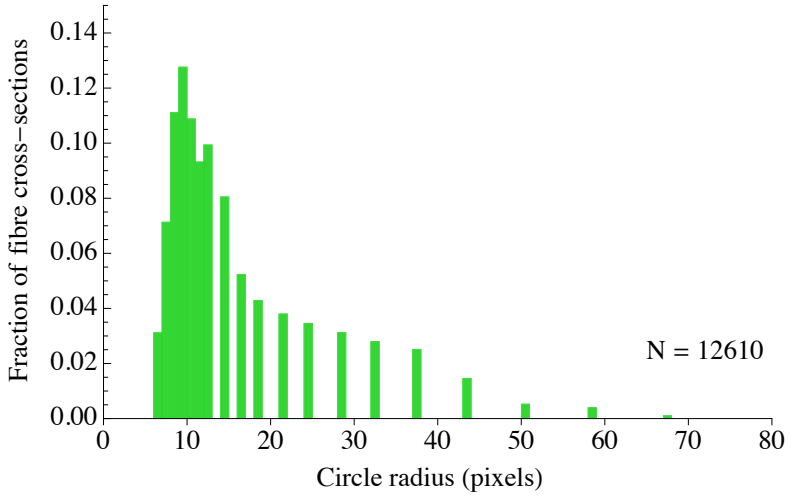


Figure 6.2: Distribution of radii for the circles best representing fibre cross-sections.

We found that picking our training sample to have a matching distribution of radii resulted in improved fibre finding performance.

Sample bias 2: Representing all overlaps approximately equally

The target output for our regression function is the predicted *ground truth overlap* for each circle. We are only interested in the cross-sections of parallel fibres and other axons that travel within ∼45° of the image plane (collectively *fibres*). Only around 1/3 of the pixels in our EM images belong to *fibres*, so at many pixels no candidate circles will overlap with any ground truth fibre cross-section. Additionally, we match circles and ground truth polygons on a one-to-one basis when calculating *ground truth overlap*. Therefore candidate circles that are not approximately centred on a single ground truth polygon and of approximately the same size will tend to have a low overlap. Taken together, this means that most of the ∼19 million potential training circles have a low overlap. Additionally, only 1.4% of the best-fitting true fibre circles for our data have an overlap greater than 0.9, so the proportion of candidate training

circles with a very high overlap will be extremely small. The fact that the distribution of target overlap values across our pool of potential training circles is so non-uniform means that our training data is *unbalanced*. The question of how to handle *unbalanced* data sets in classification and regression problems has been widely studied, and the consensus seems to favour *balancing* the training data to improve prediction performance. To do this, we pick our training sample to ensure a relatively uniform distribution of target overlap. Due to the extremely small proportion of candidate circles with very high overlaps, a trade off must be made between the uniformity of overlap achieved and the maximum supported total sample size. This is because the number of candidate circles in each overlap bin is effectively limited by the maximum number of candidates in the smallest bin. We found that a strictly uniform sample across bins with an overlap interval of 0.1 resulted in very small total sample sizes. We therefore pooled the bins with overlap greater than 0.7 into a single "high overlap" bin, which permitted a much greater total sample size.

We considered balancing the training data in terms of target overlap more important than mirroring the distribution of true fibre circle radii. Therefore we combined the two sampling strategies discussed above by first taking the maximally supported sample that mirrored the true fibre circle radii distribution and then sampling from this reduced set of candidate circles to ensure an approximately uniform distribution of target overlaps. This resulted in a supported total sample size of $\sim$65,000 from the $\sim$19 million possible circles in a training data set.

6.3 Selecting a "fibreness" feature vector

An important choice to make for our algorithm is the form of the "fibreness" feature vector that we use to encode the evidence provided by the image for each candidate fibre circle. We have chosen a feature vector based on annular BIF histograms (section 5.4). However, there are several choices to be made regarding the exact form of these histograms. Here we determine the optimal choice of BIF and annulus parameters, as well as the best BIF type and normalisation method.

6.3.1 BIF and annulus parameters

In section 5.4 we introduce a circle-based model for predicting *ground truth overlap* using annular BIF histograms as image features. Here we discuss the BIF and annulus parameters associated with this scheme in more detail and determine the optimum parameters for most accurately predicting *ground truth overlap volume*. To assess how sensitive the performance of our algorithm is to these parameters, we explored a range of values for each of them. For each parameter set, a *logistic* regression function (section 6.4.2) was trained. Training circles were

sampled as discussed in 6.2 from a training data set (id:2). Performance was compared using the root mean squared (RMS) error of the predicted *ground truth overlap* values for a *uniform* sample of circles from a separate validation data set (id:4). Figure 6.3 shows how the RMS error for our validation sample changes as we vary each of the BIF histogram parameters while holding the remaining parameters at their optimal values. The BIFs associated with a range of scales (σ) and "flat" thresholds (γ) are also visualised in figure 6.4 to provide an intuition of their effect on algorithm performance.

Gaussian scale (σ)

This is the scale of the Gaussian used as the basis for the family of Derivative of Gaussian filters used to generate BIF responses. The larger σ, the larger the features in the image must be to produce a strong response for the corresponding BIF classes. For our annular BIF histogram scheme, we expect the key information for discriminating fibres to be associated with the presence of external membrane. Therefore we expect the optimal σ to reflect the characteristic membrane thickness. For our data set this is $\sim$20 nm ($\sim$2 pixels). Visually inspecting the BIFs for a range of σ values in figure 6.4, it appears that setting σ to approximately 2 results in the clearest membrane signal. Figure 6.3a shows the effect on the regression error of varying σ. As expected, performance is sensitive to σ, with the optimal value being 1.75, close to the expected value.

"Flat" threshold (γ)

This is the response threshold below which the pixel is assigned to the *flat* (pink) BIF class. Visually inspecting the BIFs for a range of γ values in figure 6.4, it is clear that setting γ too high results in a loss of membrane information. However, it is less easy to intuit whether a low γ is better than a γ of zero. Figure 6.3b shows the effect on the regression error of varying γ. For low values of γ, performance is not sensitive to its exact value, although performance falls slightly for very low values. However, performance rapidly deteriorates as γ is increased above 0.2. This tallies with our intuition. The optimal value of γ is 0.085, in the middle of the low sensitivity range. Inspecting the BIFs in figure 6.4 closely, it can be seen that a non-zero γ has a tendency to "flatten" out some interior membrane that would otherwise be represented by dark line BIFs. This may be the explanation why the optimal γ is non-zero. However, as our annulus scheme excludes much of the fibre interior, we would expect any effect driven by this to be small. Indeed figure 6.3b shows that the impact of setting γ below its optimal value of 0.085 is much less than the impact of setting it above its optimal value.

Annulus inner radius multiplier (*$mInner$*)

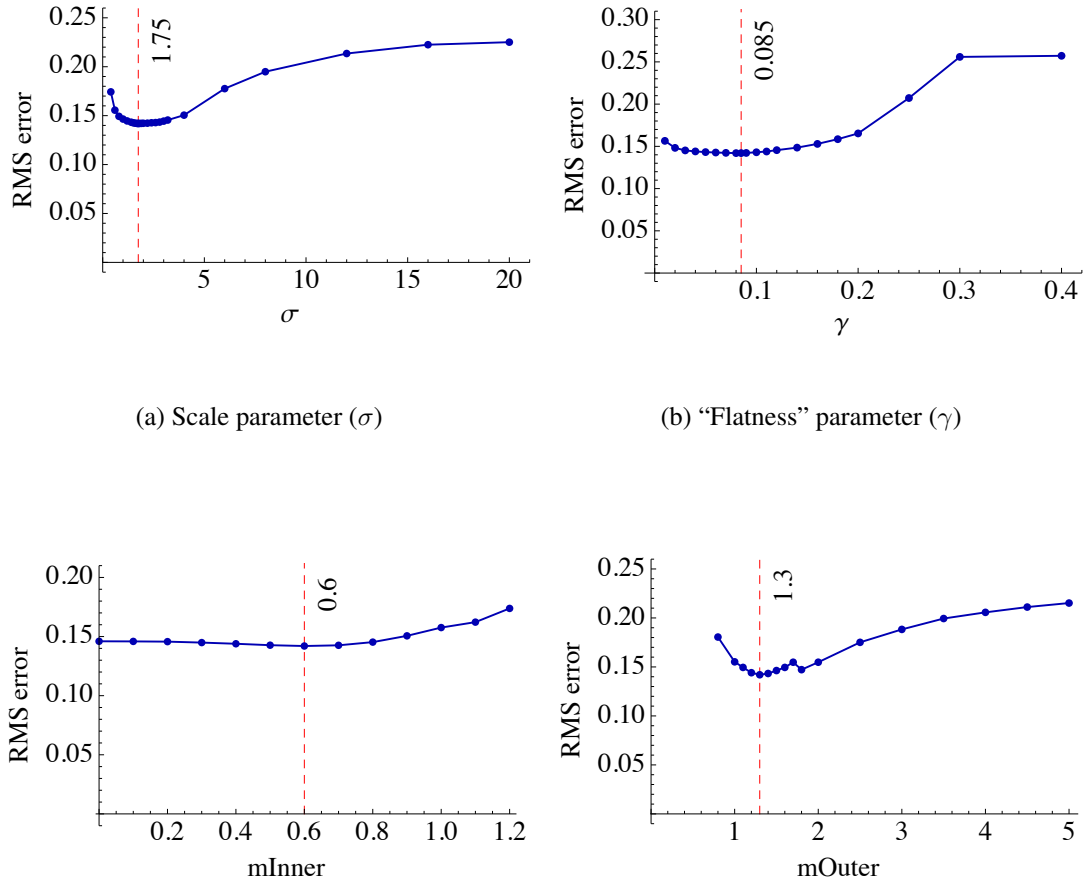
This is the multiplier mapping the radius of a candidate circle to the inner radius of the associated annulus over which the BIF histogram is calculated. Setting the inner radius to zero has the effect of including all BIFs associated with the fibre interior in the histogram. Fibre interiors can be quite cluttered with a range of randomly oriented BIFs associated with interior structure such as synaptic vesicles and mitochondria (see figure 5.8). It might therefore be reasonable to expect performance to be poorer if this interior "noise" is included. However, fibre boundaries are not perfectly circular and setting this multiplier too close to 1 might result in the exclusion of a portion of the fibre external membrane and therefore also be expected to result in poorer performance. These two considerations might lead us to intuitively expect an optimal inner radius multiplier between 0 and 1. However, examining figure 6.3c, it appears that performance is not very sensitive to this parameter, although we can identify an optimal value of 0.6.

Annulus outer radius multiplier (*$mOuter$*)

This is the multiplier mapping the radius of a candidate circle to the outer radius of the associated annulus over which the BIF histogram is calculated. Setting the outer radius to a value greater than 1 has the effect of including some BIFs associated with neighbouring fibre or non-fibre regions in the histogram. As the BIFs in these regions are not expected to be coherently oriented with respect to the fibre annulus, it might be reasonable to expect performance to be poorer if this exterior "noise" is included by setting this multiplier too high above 1. However, fibre boundaries are not perfectly circular and setting this multiplier too close to 1 might result in the exclusion of a portion of the fibre external membrane and therefore also be expected to result in poorer performance. These two considerations might lead us to intuitively expect an optimal outer radius multiplier that is greater than 1 but not by too much. Examining figure 6.3d we see the expected sensitivity of performance to this parameter, with a clear optimal value of 1.3.

6.3.2 BIF type and histogram normalisation

In 5.4.1 we introduced three types of BIF: *unoriented BIFs* (BIFs), *oriented BIFs* (oBIFs) and *radially normalised oBIFs* (rBIFs). We also discussed square-rooting histograms to make their elements have equal precision. Finally, we introduced the concept of dividing the annulus associated with a fibre circle into eight segments and taking the mean of these eight segment histograms rather than a single whole-annulus histogram. The motivation for this segment-based approach was to enforce a requirement for consistent membrane evidence along the perimeter of each putative fibre. Note that there is no difference in the mean segment-based histogram



(a) Scale parameter (σ)

(b) "Flatness" parameter (γ)

(c) Annulus inner radius multiplier

(d) Annulus outer radius multiplier

Figure 6.3: BIF histogram parameter sensitivity. The BIF scale (σ) and the multiplier for the outer annulus radius ($mOuter$) have clear optimal values. Performance is less sensitive to the BIF "flat" threshold (γ) and the multiplier for the inner annulus radius ($mInner$), though performance does significantly deteriorate if γ is set too high. Plots show the root mean squared (RMS) error between the *ground truth overlap* predicted by regression and the true *ground truth overlap* for our validation sample. Each plot shows the effect of varying a single parameter while the other three are held at their optimal values. Red dashed lines indicate optimal parameter values. The maximum value for $mInner$ is constrained to be lower than the optimal value for $mOuter$.

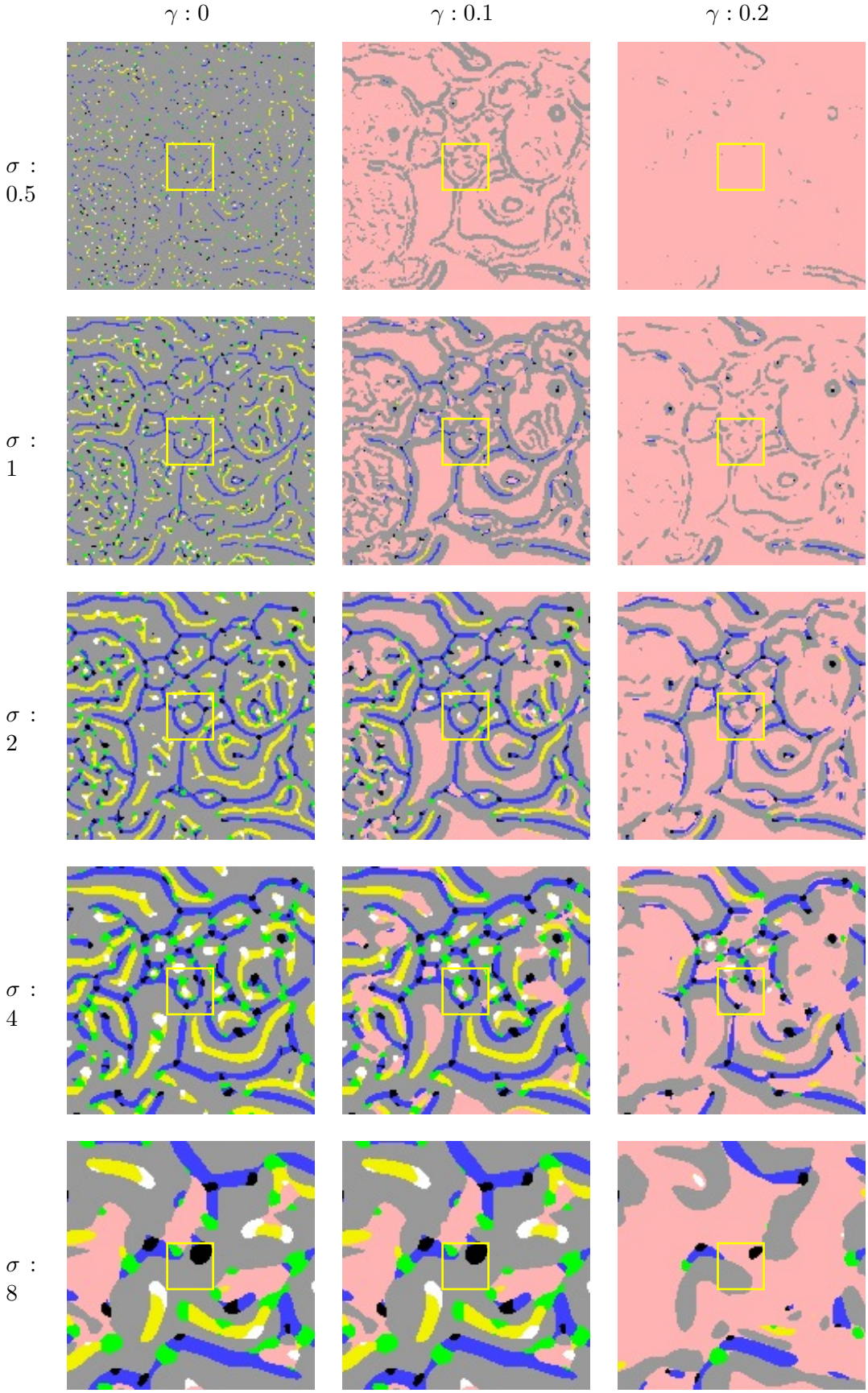


Figure 6.4: Visualising the effect of varying the scale (σ) and "flatness" (γ) BIF parameters. The optimal parameters for our annular BIF histogram scheme are $\sigma = 1.75$ and $\gamma = 0.085$ (see figures 5.8 and 6.3). The yellow box marks the fibre neighbourhood in figures 5.9 and 5.10. The central image (σ=2; γ=0.1) shows close to optimal parameters. When σ gets too small (0.5) or too large (8), the information required to reliably distinguish fibres is lost.

and the single whole-annulus histogram unless the histograms are square-rooted.

To determine the effect of BIF type and histogram normalisation method on the performance of our algorithm, we compared the end-to-end fibre finding performance for all six permutations of BIF type and normalisation method. After training a linear regressor using a sample from a training data set (id:2) using algorithm 3, we predicted the *ground truth overlap volume* for an independent test set (id:4) using algorithm 4 and found circles using algorithm 5. For this initial investigation we did not tune a stopping threshold for algorithm 5 on a separate data set. Instead we placed the same number of circles as there were true ground truth polygons. By definition *precision*, *recall* and *f-measure* are all equal when the number of found circles equals the number of true fibres. While the maximum *f-measure* is usually achieved by placing more or less circles than this, the *f-measure* achieved by placing the expected number of circles is usually quite close to the maximum achievable. We are likely to somewhat overestimate the *f-measure* achievable on unseen data using this approach. However, comparisons across different BIF types and normalisation methods will be valid as the same stopping criteria is used for all conditions. The algorithm also makes use of a fixed luminance threshold for mitochondria exclusion. This was set globally to 0.24 for all data sets across all experiments and so will also not effect the validity of the comparison.

Figure 6.5 compares the *overlap f-measure* achieved on the test data set (id:4) for all six combinations of BIF type and histogram normalisation method. It can be seen that square-rooting histograms to equalise the precision of their elements results in improved performance for all BIF types. The performance benefit associated with square-rooting histograms is of the same order as that gained by moving from BIFs to rBIFs and it seems clear that square-rooted rBIF histograms perform best overall.

6.3.3 Adding non-BIF image features

We explored adding non-BIF features such as luminance, circle radius and even the membrane probability map produced by an alternative pixel-based classifier (*ilastik*: Sommer et al., 2011). However, these additional features had little effect on performance, even when combined with a flexible random forest classifier (see *Forest-All* in figure 6.6).

6.4 Selecting a method to predict ground truth overlap

One of the key choices to make for our algorithm is the mapping we learn to transform our annular BIF histograms into predictions of *ground truth overlap*. The general term for learning a mapping from an N-dimensional vector input $\boldsymbol{x}$ to a scalar output y is *multiple regression*. In our final scheme, the input $\boldsymbol{x}$ is the square-rooted annular rBIF histogram associated with a

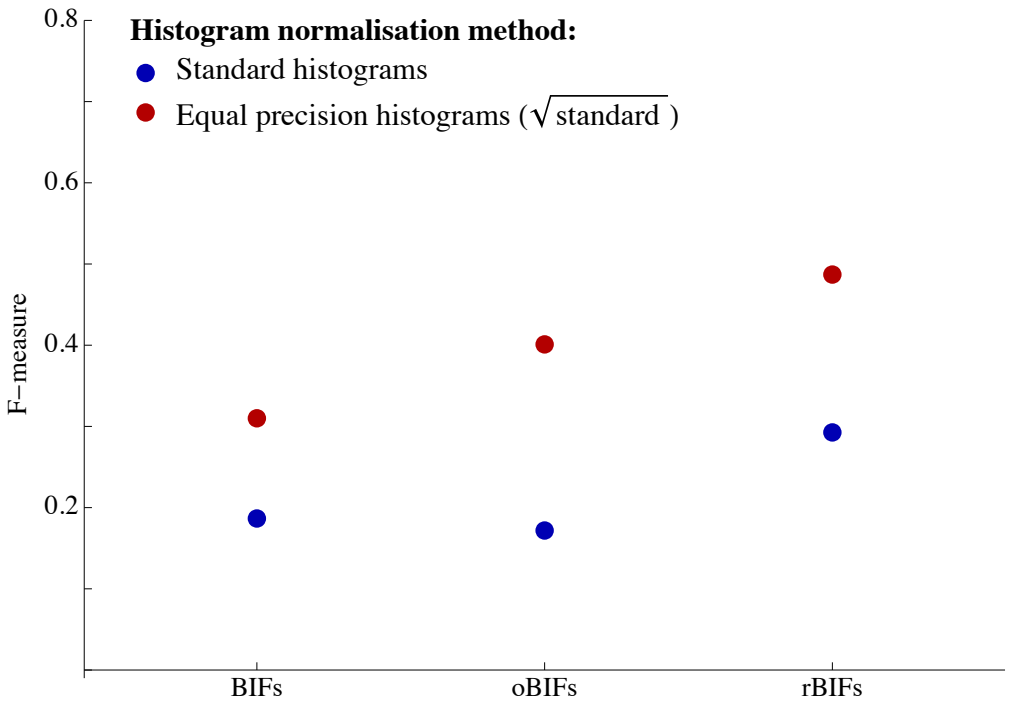


Figure 6.5: BIF type and normalisation. Square-rooting histograms to equalise the precision of their elements results in improved performance for all BIF types.

circle. The output y is the predicted *ground truth overlap* for the circle. We explore a range of regression methods, from simple *linear regression* to very non-linear *random forest* regression. We also explore combining *classification* and *regression* methods.

To determine the effect of regression method on the performance of our algorithm, we compared the end-to-end fibre finding performance for each explored method. After training a regression function using a sample from a training data set (id:2) using algorithm 3, we predicted the *ground truth overlap volume* for an independent test set (id:4) using algorithm 4 and found circles using algorithm 5. For this initial investigation we did not tune a stopping threshold for algorithm 5 on a separate data set. Instead we placed the same number of circles as there were true ground truth polygons. The algorithm also makes use of a fixed luminance threshold for mitochondria exclusion. The use of a fixed stopping and mitochondria exclusion criteria are discussed in more detail in section 6.3.2.

6.4.1 Linear regression

Linear regression generates a scalar output y from an N-dimensional vector input $\mathbf{x}$ by taking a linear weighted sum of the vector components and adding a fixed offset (see eq. 6.1).

$$y = a + \mathbf{b} \cdot \mathbf{x} = a + \sum_{i=1}^{N} b_i x_i \tag{6.1}$$

The optimal weight vector $\mathbf{b}$ and offset a can be determined by minimising the total squared difference between the desired target outputs and the outputs predicted from the regression across a set of training samples. This process is known as *least-squares* fitting and the optimal $\mathbf{b}$ and a can be directly calculated from the target output values and the input vectors for the training samples. To determine linear regression weights and apply them to new data we use the Matlab functions *glmfit* and *glmval*, with *distribution* set to *normal* and *link* set to *identity*.

6.4.2 Logistic regression

When the regression output is restricted to lie between 0 and 1, as *ground truth overlap* is, logistic regression can be a more appropriate choice of mapping. In logistic regression a linear weighted sum of the input $\mathbf{x}$ is passed through the *logistic* function $f(t) = \frac{1}{1-e^{-t}}$ in order to generate the predicted output. Equivalently, the target output y can be transformed via the *logit* function $\tilde{y} = ln(\frac{y}{1-y})$ and linear regression performed on the transformed output (see eq. 6.2).

$$y = \frac{1}{1+e^{-(a+\mathbf{b}\cdot\mathbf{x})}} = \frac{1}{1+e^{-a}\prod_{i=1}^{N} e^{-b_i x_i}} \rightarrow \tilde{y} = ln\left(\frac{y}{1-y}\right) = a + \mathbf{b}\cdot\mathbf{x} = a + \sum_{i=1}^{N} b_i x_i \tag{6.2}$$

To determine logistic regression weights and apply them to new data we use the Matlab functions *glmfit* and *glmval*, with *distribution* set to *binomial* and *link* set to *logit*.

6.4.3 Random forest regression

Binary decision trees have been used for regression problems since at least the mid 1980s (Breiman et al., 1984). During the construction of such a tree, the training data is recursively split using a series of binary decision functions to minimise some measure of the *impurity* of the two subsets produced by the split. In theory these decision functions can be very sophisticated. However, in practice the decision function is usually a simple threshold on a single input variable (an axis-aligned linear split). At each split, the input variable and threshold are selected to minimise the impurity of the subsets. For regression, the impurity measure is usually the *sum of squared errors* of the target values across the two subsets (i.e. $\sum_{i \in set_A}(y_i - \bar{y}_A)^2 + \sum_{j \in set_B}(y_j - \bar{y}_B)^2$) or an equivalent measure such as the standard deviation of the target output across the two subsets. If the impurity of the subsets is the same as the parent set then the node is not split and becomes a *terminal* or *leaf* node. Node splitting can also be halted based on other stopping criteria such as minimum node population or maximum tree depth. For regression, each terminal node is associated with a *value function* used to infer

the output for unseen input data assigned to the node. Often this is simply the mean of the target values assigned to the node during training. Once trained, each new data point is processed by the series of binary decision functions and is assigned to a terminal node. If the value function for the node is simply the mean training target values, a binary decision tree can be considered as approximating a k-nearest neighbour algorithm where k may vary between terminal nodes.

A single binary decision tree is difficult to optimise. If the tree does not grow enough, it will not predict the training data well. However, if the tree is permitted to grow too deep (or terminal nodes grow too small) the tree will overfit the training data and not generalise well to unseen data. Many methods have been explored to find the optimal decision tree including pruning. A *random forest* of decision trees is an ensemble method that addresses overfitting by combining multiple decision trees. Rather than attempting to optimise each tree, properties of the trees are randomly varied and the mean prediction across the trees in the forest is taken as the output. Ensemble methods have shown themselves capable of combining many weak predictors into a strong predictor if the individual weak predictors are not strongly correlated. By randomly varying key properties of its constituent decision trees, a random forest ensures that they are not strongly correlated and benefit from being combined in an ensemble. Originally proposed by Breiman (2001), random forests combine two existing approaches to randomising decision trees: *bagging* (Breiman, 1996) and *random subspace selection* (Ho, 1998). Bagging involves taking a random sample of training data for each tree. Usually, these samples are the same size as the training data, but are drawn *with replacement*. This means that some training data may be selected multiple times and some might not be selected at all. On average each sample will contain approximately 63% of the training data, meaning each tree is trained on different data drawn from the same distribution. Thus, while individual trees may overfit their subset of the training data, the ensemble will not overfit the training data as a whole. The fact that there is a subset of approximately 37% of the data that is not used to train each tree also permits an estimate of the generalisation error for the ensemble to be made using this *out of bag* data. Subspace selection involves considering only a random subset of the input variables when determining the decision function at each node. This further decorrelates the trees, making the ensemble approach more effective. In this work we use Jaiantilal's MATLAB wrapper to Liaw and Wiener's C code provided as part of the R *randomForest* package (MATLAB: Jaiantilal, 2012; C: Liaw and Wiener, 2002).

6.4.4 Comparison of regression methods

Figure 6.6 compares the *overlap f-measure* achieved on the test data set (id:4) for *linear*, *logistic* and *random forest* regression methods. Linear regression achieves an *overlap f-measure* of 0.49

using square rooted histograms. Logistic regression performs only slightly better than linear regression for square-rooted histograms, and random forest classifiers using only BIF features perform slightly worse. The small differences in the performance observed using square-rooted histograms is within the variation in performance across data sets. Therefore there does not appear to be a significant difference in performance between regression methods across data sets when using square-rooted histograms. However, the random forest regression performs much better than linear or logistic regression when non square-rooted histograms are used. This, coupled with the fact that random forest performance does not improve when the number of trees in the forest is doubled, suggests that 100 trees is sufficient to learn the full structure present in the data (*Forest-100* vs. *Forest-200* in figure 6.6). Normalising BIF histograms to have equal precision elements seems to be one of those "simple things that work", providing a similar improvement to using a much more powerful learner. In fact, adding additional image features such as image luminance, circle radius and even the output of an alternative pixel classifier (*ilastik*: Sommer et al., 2011) into the random forest regression does not appear to significantly improve peformance beyond that provided by square-rooting the BIF histograms (*Forest-All* in figure 6.6).

We selected *logistic* regression as our preferred method as it achieved the highest *f-measure* while also being much faster to train and run than a random forest.

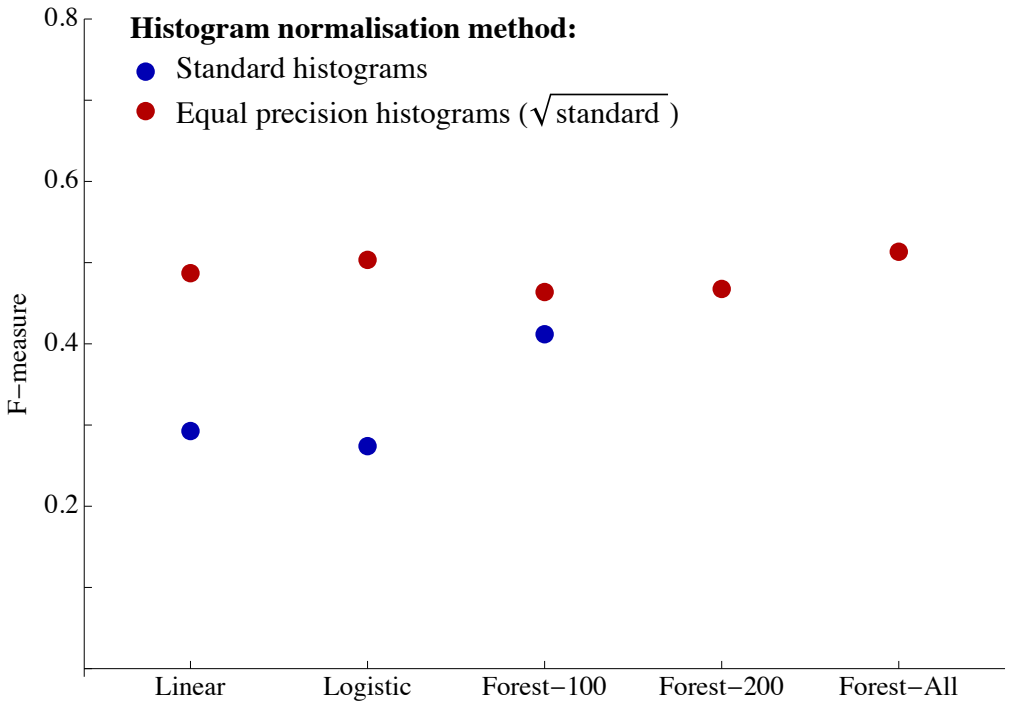


Figure 6.6: Regression method. Using a powerful random forest for regression improves performance when standard BIF histograms are used. However, logistic regression achieves the best performance when equal precision histograms are used.

6.4.5 Combining classification and regression

While we have demonstrated that accurately predicting *ground truth overlap* is *sufficient* for finding suitable circles, it is clear that it is not *necessary*. In the case where we predict an overlap of 1 for each of the best fitting circles and an overlap of 0 for all other circles, we will certainly find these best-fitting circles. More generally, it is important that the best fitting circles score higher than the circles we don't want to find, but it is less important that we accurately predict the overlap for the circles we do not want to find. This intuition is supported by the fact that over-representing high overlap circles in our training data improves the performance of our algorithm (see section 6.2). It is therefore reasonable to ask whether learning to accurately predict overlap for circles is the best approach. One alternative is to initially classify circles as either *low* or *high* overlap circles. Low overlap circles can safely be ignored and training a second regression stage on the remaining high overlap circles might be expected to result in improved overlap prediction accuracy.

To explore this, we combine the random forest regression approach described in section 6.4.3 with an initial random forest classifier to distinguish between low and high overlap circles. A random forest classifier is very similar in structure to a random forest regressor. It uses the same *bagging* and *random subspace selection* techniques to generate a strong ensemble learner. However, the structure of the binary decision trees is slightly different. Node splitting occurs in the same manner, but the *impurity* measure used to evaluate candidate splits is the *binary classification error* and the *value function* used to infer a value for an unseen data point is the *majority class* for the training data assigned to a terminal node.

We compared the performance of this two-stage classification and regression approach against a pure regression approach using the same method of evaluating *overlap f-measure* on a test data set (id:4). Figure 6.7 compares the performance of a single-stage *logistic* regression against four alternative two-stage models. Each of the two-stage models varies only in the overlap threshold used to assign circles to the low or high overlap groups when training the initial classification stage. It can be seen that there is no significant difference in performance between any of the two-stage classifiers and a single stage *logistic* regression.

6.5 Algorithm performance

6.5.1 Visualising algorithm performance

Figure 6.8 shows the fibre circles found by our algorithm on one of the test data sets (id:8), along with the ground truth fibre polygons for the data set. The labelling produced by our circle-based algorithm is sparse, reflecting the sparse nature of the ground truth quite well. The

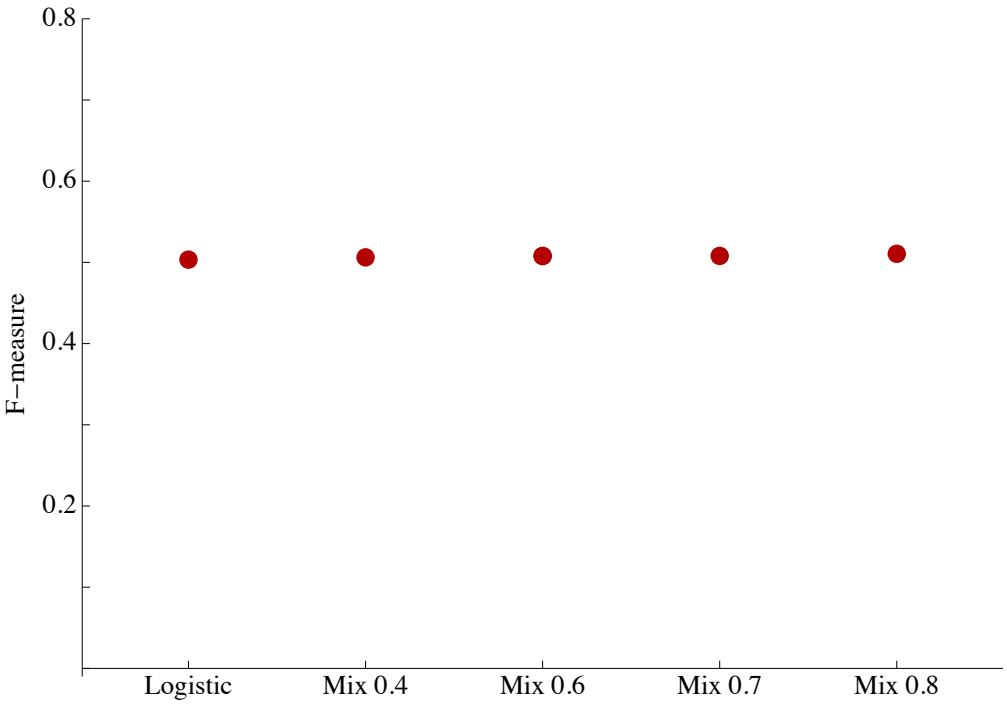
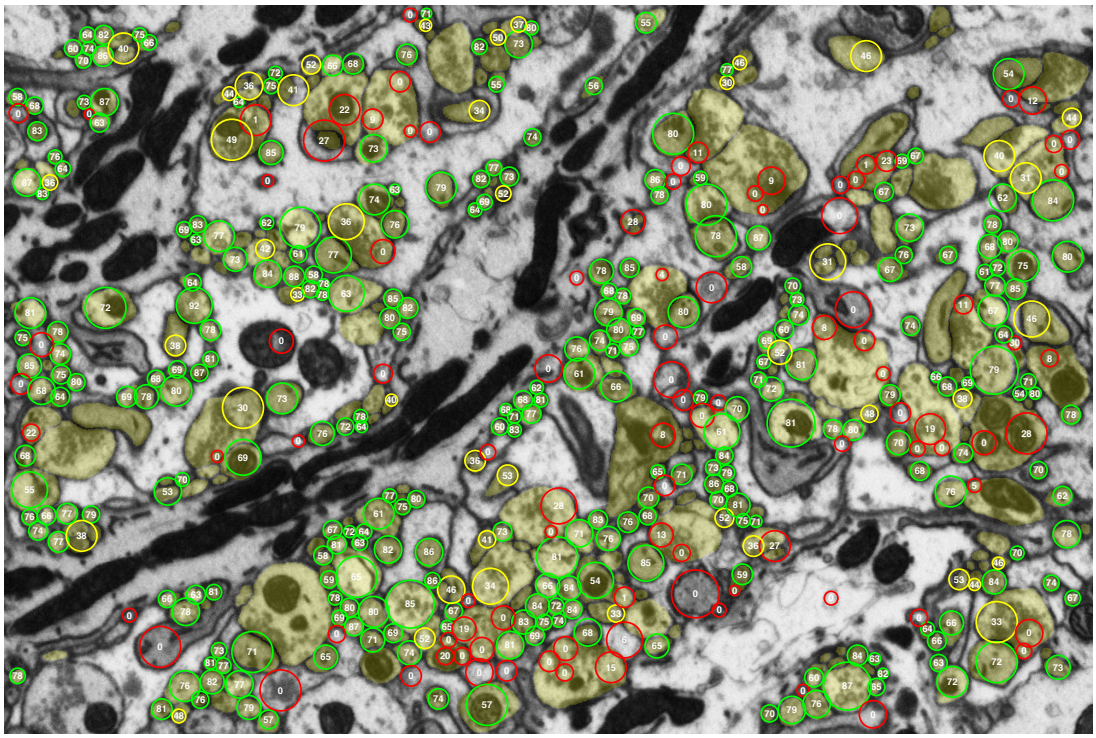


Figure 6.7: Combining classification and regression. Using a two stage approach that first classifies circles as high or low overlap and then performs regression for high overlap circles does not improve performance.

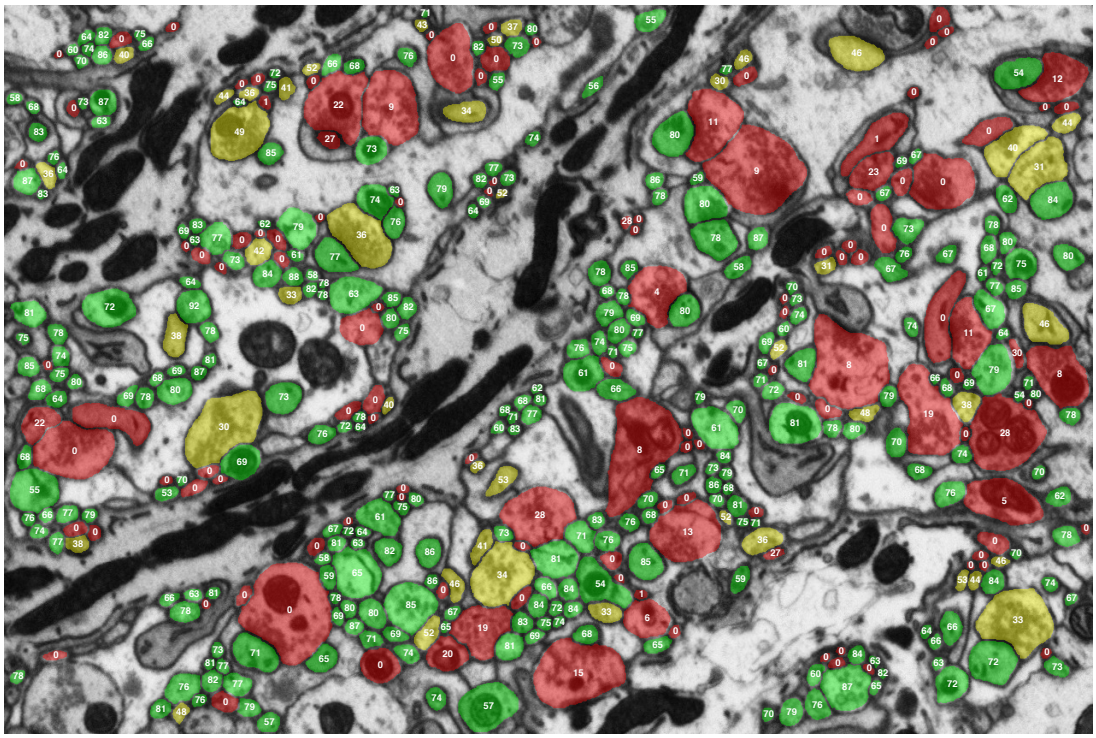
found circles are predominantly constrained to the areas of the image where the ground truth fibres exist, with few circles found in areas where there are no fibres. For the example data set shown, only 9.4% of the found circles (40/424) are placed completely outside of the ground truth fibres, despite this "non-fibre" area accounting for 66% of the image. Sometimes multiple circles are placed within a single true fibre polygon. As the algorithm scoring only permits one found circle to be associated with each true fibre polygon, only the circle that most overlaps with each ground truth fibre will contribute to the *overlap f-measure*. However, permitting multiple circles to be matched with a each true fibre polygon would not be expected to improve performance significantly as most of these additional unmatched circles would have low overlap with the ground truth. It is clear the algorithm struggles to find representative circles for many of the large irregular cross-sections. These are pre-synaptic boutons, which occur when a fibre swells to make a synapse. These contain significant intracellular "clutter" such as vesicles and mitochondria. The algorithm also struggles to find some smaller fibre cross-sections. This is discussed further in section 6.5.3.

6.5.2 Analysing algorithm performance

Figure 6.9a shows the distribution of overlaps between true polygons and found fibres pooled across four test data sets (ids: 3, 5, 7, 8). Any true polygons that are not overlapped by any



(a) Fibre circles found using the circle-finding algorithm



(b) Fibre polygons from manually labelled ground truth

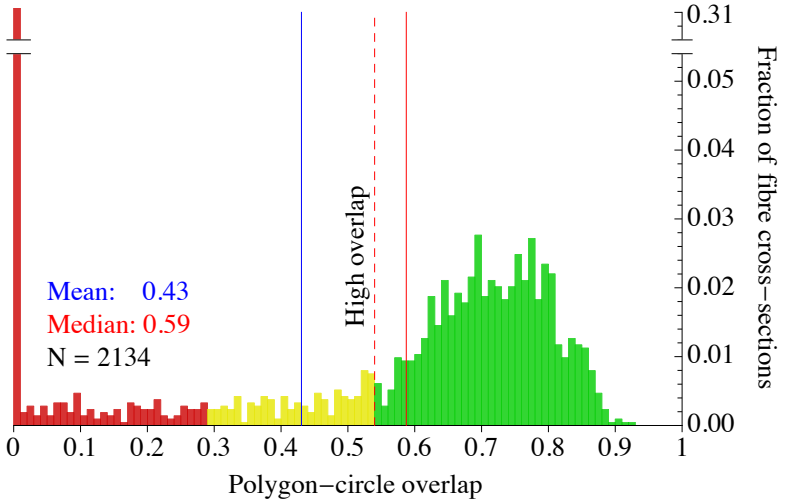
Figure 6.8: **(a)** Fibre circles found using the best algorithm parameters, along with the underlying ground truth polygons for comparison. Circle colour coding indicates *high* (green), *medium* (yellow) and *low* (red) overlap with the ground truth polygons. Actual overlap percentages are indicated by white text. **(b)** The corresponding manually labelled ground truth polygons colour coded by overlap using the same green/yellow/red scheme to indicate high/medium/low overlap. Again, actual overlap percentages are indicated by white text.

found circle contribute to the zero-overlap bin. Similarly, any found circles that do not overlap with any true polygons also contribute to this bin. The corresponding distribution for the best possible circles that can be found from perfectly predicted *ground truth overlap* is shown for comparison (6.9b, duplicate of 5.6a). The distribution of overlaps for polygon-circle pairs is clearly much wider than that for the best possible circles, with significantly lower mean and median. Across the pooled data sets 75% of placed circles and 72% of true polygons have *medium* or *high* overlap. While these are relatively high percentages, they compare poorly to the corresponding 97% and 100% percentages for the best possible circles. This is also reflected in the overall pooled *overlap f-measure* across the four test data sets. At 0.51, this is significantly lower than the 0.77 achieved for the best possible circles. In section 6.5.3 we examine the poorly found fibres more closely and in section 6.7.5 we use the *overlap f-measure* to compare the performance of our algorithm against that of an alternative pixel-based segmentation approach.

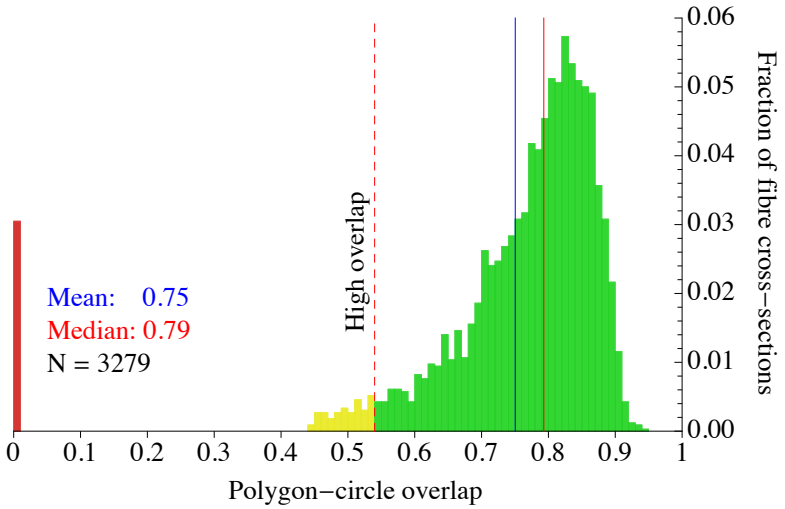
6.5.3 Properties of poorly found fibres

The algorithm seems to find most small, round fibres well but struggles to find larger, irregular fibres. It also seems to miss some very small fibres. Many of the poorly found fibre cross-sections are pre-synaptic boutons. These are points where a fibre swells and makes a synapse, and their cross-sections are both large and irregular. They also contain significant intracellular membrane "clutter" in the form of synaptic vesicles and mitochondria. Figure 6.10 shows how the distribution of overlap scores varies with the size and eccentricity of the true fibres (data pooled across data sets 3, 5, 7 and 8). The algorithm fails to find fibre cross-sections well if they are very small (<188 pixels), very large (>1397 pixels) or too irregular in shape (eccentricity>1.4). The poor performance on very small and very large fibres remains when only very circular fibres are considered (data not shown).

As many of the larger poorly found cross-sections are boutons, a key question is whether the poor performance for boutons is due to their irregular shape or their associated intracellular clutter. An irregular membrane might result in a wider distribution of radially normalised orientations for corresponding rBIFs, which might make high and low overlap rBIF histograms less easy to distinguish. Intracellular membrane "clutter" might be expected to contribute a spurious membrane signal to circles that are not well representative of fibres. In fact mitochondria alone can often provide sufficient coherently oriented membrane signal to result in the erroneous placement of a circle. The addition of an image luminance threshold for mitochondria exclusion substantially mitigates this problem, but does not completely eliminate it (section 6.1.2). Most remaining poorly placed circles appear to be the result of the accumulation of coherently oriented membrane signal across a collection of vesicles, often combined with ad-



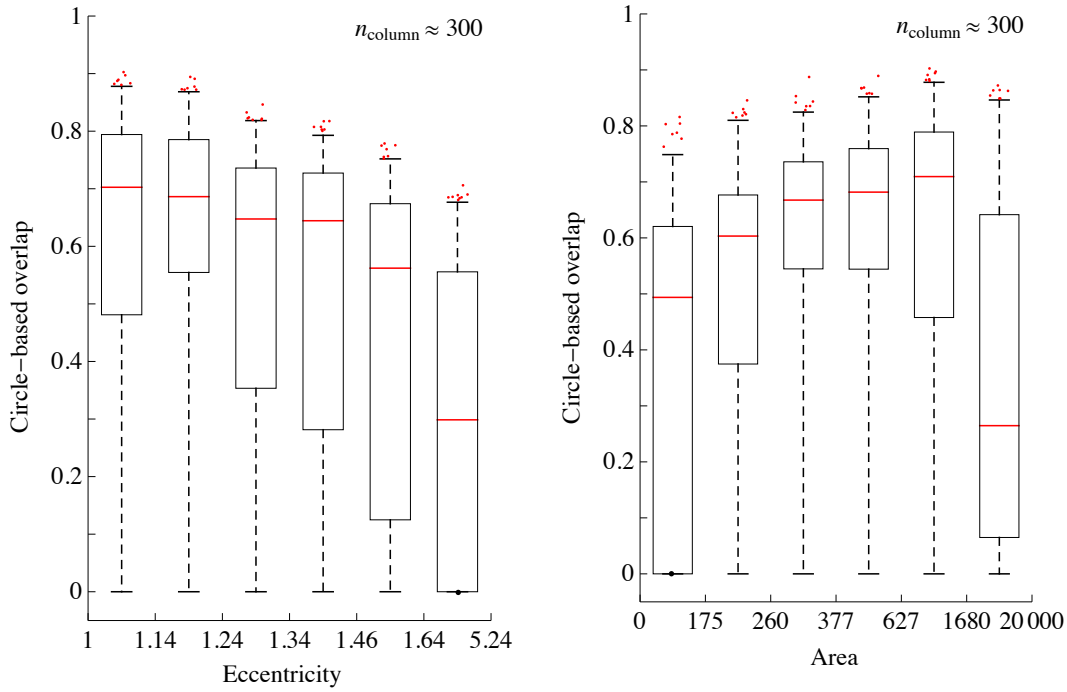
(a) Distribution of ground truth overlap for circles found by our algorithm



(b) Distribution for circles found from the *ground truth overlap* volume

Figure 6.9: Comparing overlap distributions for **(a)** circles found by our algorithm and **(b)** circles found from the ground truth *ground truth overlap volume* (duplicate of figure 5.6a). Although 75% of algorithm-found circles have high or medium overlap with true fibre polygons, this is significantly worse than for the best possible circles that can be found from perfectly predicted *ground truth overlap* (97%). The zero-overlap bin (far left) in **(a)** includes both unpaired found circles (13.5%) and unpaired true polygons (17.5%). Note the broken axis to accommodate the height of this bin. There were no unpaired found circles in **(b)**. Bar colours indicate *low* (red), *medium* (yellow) and *high* (green) overlap. Mean and median overlaps are indicated with blue and red solid vertical lines. Data pooled across data sets 3, 5, 7 and 8.

ditional membrane signal from either mitochondria or external fibre membrane. This suggests that the presence of intracellular membrane plays a significant role in reducing fibre finding performance.



(a) Overlap vs eccentricity

(b) Overlap vs area

Figure 6.10: **(a)** The overlap achieved by our algorithm as a function of the *eccentricity* of true fibres. **(b)** The overlap achieved by our algorithm as a function of the *area* of true fibres. Box bounds indicate 25th and 75th percentiles. Whisker bounds indicate 2.5th and 97.5th percentiles. Note that an approximately equal number of data points ($\sim$ 300) contribute to each box-whisker column, resulting in non-uniform column width on the x-axis. Data is pooled across data sets 3, 5, 7 and 8. Circles were found using a regressor trained using data set 1 and an overlap stopping threshold tuned on data set 4.

6.6 Comparing reconstruction methods across studies

In this section we consider the difficulties of comparing reconstruction methods across studies and explain why existing benchmark data sets for EM reconstruction are not suitable to enable reliable cross-study comparison for our task.

6.6.1 Issues when comparing results across studies

There are several factors that make inter-study comparison difficult for methods of reconstructing neurons from EM images. Some of these issues, such as choice of similarity measure and density of reconstruction, can even make comparison between techniques within a single study difficult (see section 6.7.3).

Tissue staining

There are two main types of staining used for making EM images of neural tissue. The first is the classical *intracellular* stain. This stains all interior and exterior cellular membrane with heavy metals, making it appear dark in the images. This type of staining makes details of intracellular structures visible, including the vesicles and post-synaptic densities (PSDs) that identify synaptic connections between neurons. However, as both interior and exterior membrane are stained the same, they are difficult to distinguish from each other and the presence of intracellular "clutter" makes the problem of reliably segmenting neural cross-sections difficult. The second type of stain is an *extracellular* stain, which only stains exterior membrane. The absence of intracellular "clutter" makes the segmentation task significantly easier. However, in most types of neural tissue the lack of intracellular staining of vesicles and PSDs makes it much more difficult to unambiguously identify connections between neurons. It is likely that most segmentation approaches will perform significantly better on data sets that use the extracellular stain. Therefore fairly comparing reconstruction methods across studies that use different stains is essentially impossible.

Imaging method

There are two imaging factors that can affect the performance of segmentation methods: (i) whether slicing occurs before or after imaging (serial section vs. serial block face microscopy) and (ii) whether imaging uses transmitted or reflected electrons (transmission vs. scanning microscopy). Imaging methods that remove a slice from the sample prior to imaging (serial section imaging) are limited in their minimum achievable z-resolution as the slice must remain thick enough to handle. This typically limits z-resolution to $\sim$50 nm, although some labs are pushing this as low as 30 nm using automated collection methods (Hayworth et al., 2007; Schalek et al., 2012). Given that axons can shrink to $\sim$50 nm in diameter in places, this limited z-resolution can make thin axons running at an angle to the slice plane very hard to track.

For methods that image using electrons that have passed through the slice (transmission imaging), cell membranes can also be significantly blurred for neurites running at angles to the slice plane due to the spatial averaging effect of the resulting maximum projection image. For methods that image using reflected neurons (scanning imaging), this blurring effect can be eliminated. However, the thickness of serial section slices can still make it difficult to reliably match cross-sections of neurites from slice to slice if they run at an angle to the slice plane. The delicate nature of the tissue slices used in serial section imaging also means that slices are regularly distorted, damaged or even entirely lost prior to imaging. While a whole research field exists to reliably register these slices together post imaging (Anderson et al., 2009), this

problem effectively disappears for methods that image each slice prior to removal (serial block face imaging). Block face imaging methods also permit much thinner z-slices to be removed between images (∼10 nm using focussed ion beam milling), reducing many of the tracking issues for thin neurites and those running at an angle to the slice plane. However, block face methods are limited to imaging using reflected electrons and therefore often have higher noise, lower contrast and lower within-slice resolution than transmission images obtained via serial section methods. These differences in image properties make fairly comparing reconstruction methods across studies using different imaging methods difficult.

Tissue type

Different regions of the brain can have very different structures. The most obvious difference is that between grey and white matter. Axons in grey matter are separated by thin membrane boundaries and generally run in all directions, while axons in white matter are separated by much thicker myelinated boundaries and tend to run parallel to other fibres for significant distances in axon bundles. However, even within grey matter there can be significant differences in the geometry of neurites. For example, in the retina synapses can be reliably identified from contact geometry alone (Briggman, Helmstaedter, and Denk, 2011), permitting the use of extracellular staining without sacrificing information about connectivity.

In this work we examine the molecular layer of the cerebellum, which is dominated by parallel fibres (granule cell axons) running in approximately the same direction for large distances. We focus on the restricted problem of tracking these non-branching parallel fibres. This is an important problem as these fibres represent the sole output of the cerebellar granule cells, which comprise ∼80% of the neurons in the brain (Azevedo et al., 2009). It is also a difficult problem as these fibres are some of the longest unmyelinated neurites in the brain, running for millimetres with thin 20 nm membranes and shrinking in size to as small as 50 nm in diameter in places. However, as these fibres are oriented approximately parallel with one another, we can image perpendicular to the prevailing fibre orientation so that most fibre cross-sections will be well-represented within a 2D image. This allows the application of a simpler and more computationally efficient 2+1D reconstruction approach. It also reduces any issues associated with membrane being blurred due to the limited z-resolution provided by some imaging methods (although this final benefit is not applicable to our data set, due to our high isotropic resolution). Given the differences between structure that can exist between tissue types, reliably comparing reconstruction methods across studies using different tissue types can be difficult.

Similarity measure

Many studies measure reconstruction performance using *binary pixel classification accuracy* to report the proportion of pixels successfully classified as *membrane* or *neurite interior*. As discussed in 3.5, this measure is usually not reflective of object-level segmentation quality. There has been a welcome trend towards reporting the more suitable *Rand index*, which measures pixel pair labelling accuracy and is more reflective of object-level segmentation quality. However, it is not clear how suitable this measure is for comparing sparse reconstructions such as ours that do not consist of a dense labelling of pixels. Higher level measures such as the topological *warping accuracy* have also been proposed (Jain et al., 2010), though this measure is expensive to compute and a consensus needs to be reached on the degree of warping to allow before inter-study comparisons can be reliably made. The lack of consistency in choice of similarity measure across studies makes comparison of results difficult.

Density of reconstruction

Our reconstruction goal differs from those of most studies as we aim to reconstruct only a subset of the cells within our data set. This makes it more difficult to fairly compare the results of our algorithm against results from algorithms that output a dense reconstruction of all cells. This issue is discussed in more detail in section 6.7.3.

6.6.2 Existing benchmarks and comparable results

While other computer vision research areas have standard benchmark data sets available, until recently no such data set has existed for electron microscope images of neural tissue. However, over the past two years, two data sets have been released as part of a challenge workshop at the annual International Symposium on Biomedical Imaging (ISBI). The first is a Drosophila ventral nerve cord data set with 2D binary labelling and the second a mouse cortex data set with 3D object identity labelling. These new benchmarking data sets are welcome contributions to the community and address some of the issues highlighted in section 6.6.1. Unfortunately, tissue type is still an issue for us. As we are focussed on a reconstruction problem that lends itself well to a 2+1D reconstruction approach, we have therefore developed a 2+1D algorithm. However, we cannot expect this approach to perform well on these benchmark data sets, where neural fibres can run in all directions. Therefore these are not suitable benchmarking data sets for our algorithm and we have chosen to perform our own benchmarking by evaluating an alternative state of the art reconstruction approach on our own data set (see section 6.7). We will be releasing our images and 3D object labelling to provide an additional benchmarking data set to the community.

Results for another study tracking parallel fibres in mouse cerebellar tissue have recently been reported (Jurrus et al., 2013). The similarity of the tissue, staining and imaging types minimise most of the cross-study comparison issues discussed in section 6.6.1. The underlying image and label data have not been released as a benchmark data set, and there remain some issues making a direct comparison. However, we will benchmark the performance of our 3D algorithm against that reported in this study (section 7.4).

The two ISBI benchmark data sets and the cerebellar parallel fibre data set used in the Jurrus study are briefly summarised below.

ISBI 2012 data set (2D)

This data set was released for the 2012 ISBI EM segmentation challenge. It comprises two 512×512×30 voxel volumes from classically stained Drosophila neural tissue from the ventral nerve cord of a first instar larva. These volumes were imaged at 4×4×50 nm using a serial section transmission method. For one of these volumes binary labels indicate whether pixels are membrane or neurite interior. The labelled volume is used to train and test segmentation algorithms while the remaining unlabelled volume is used to generate a segmentation for third-party evaluation via the challenge website. Performance is evaluated using three different measures: (i) the binary pixel accuracy; (ii) the maximal Rand f-measure and (iii) the minimum splits and mergers warping error. Data and documentation are available on the challenge website (Cardona, 2012) and performance on the test data can be evaluated using the challenge website (Shaar, 2012).

ISBI 2013 data set (3D)

Also known as the SNEMI3D or AC4 data set, this was released for the 2013 ISBI EM segmentation challenge. It comprises two 1024×1024×100 voxel volumes from classically stained mouse neural tissue, likely from layer 5 somatosensory cortex. These volumes were imaged using an automated serial section scanning method (ATUM: Automated Tape-collecting Ultra Microtome; Schalek et al., 2012) and have a resolution of at 6×6×30 nm. For one of these volumes neurite cross-sections are labelled with 3D object identifiers that are consistent from slice to slice for each neurite. The labelled volume is used to train and test segmentation algorithms while the remaining unlabelled volume is used to generate a segmentation for third-party evaluation via the challenge website. Performance is evaluated using the 3D object level Rand index. Data, documentation and evaluation of performance on the test data are available on the challenge website (Shaar, 2013).

Mouse cerebellar parallel fibre data set

Very recently the results of applying a sequential neural network approach to reconstruct parallel fibres in the molecular layer of mouse cerebellar tissue have been published (Jurrus et al., 2013). This data set used classical intracellular staining and a serial block face imaging method (though with coarser z-resolution than ours). It comprises a 4096×4096×400 voxel volume imaged at 10×10×50 nm resolution. Manual labelling was gathered for a 700×700×70 voxel sub-volume. From this sub-volume, 42 randomly selected z-slices were used for training while the remaining 28 were used for evaluating the 2D segmentation performance of the algorithm. The full training sub-volume was used to evaluate 3D reconstruction performance. The full 4096×4096×400 voxel image volume is available from the Cell Centered Database (CCDB) with accession number 8192 (Bushong and Deerinck, 2013), although the ground truth labelling is not available and so it is not possible to evaluate other algorithms on the same test data.

6.7 Benchmarking against ilastik on our data set

In order to benchmark our model-based approach against the state of the art, we compared its performance to that achieved by *ilastik* (Sommer et al., 2011), a recently published pixel-based random forest classifier. We chose *ilastik* because a similar classifier had been reported by the same group to perform well at finding cell boundaries in electron microscope images of neural tissue (Andres et al., 2008). Applying *ilastik* to our data set mitigates most of the issues discussed in section 6.6.1. We describe the *ilastik* reconstruction pipeline and address the issue of fairly comparing the output of the two approaches before discussing their relative performance on our data set.

6.7.1 The ilastik reconstruction pipeline

We outline the process of training the *ilastik* pixel classifier and converting its output into a 2D segmentation.

Feature selection

The version of *ilastik* used for this study (v0.5) has a range of image features available, including ones based on similar Derivative-of-Gaussian filters as those used in the BIF scheme. These features are available at a range of scales, ranging from $\sigma = 0.3$ to $\sigma = 10$. The classifier trained here uses all features at all scales and is trained on the full extra-cellular membrane pixel labelling for the training data set (id:1). As random forest classifier performance is tolerant to the inclusion of additional uninformative features, the full set of *ilastik* features was used, with no attempt made to select the most informative subset. Labelled membrane pixels are assigned to one *ilastik* class and all unlabelled pixels are assigned to a second class.

Training the pixel classifier

At each training iteration, *ilastik* outputs an estimate of the probability that each pixel is extracellular membrane. This probability estimate is thresholded, assigning pixels with estimated probability ≥ 0.5 to the membrane category and the remaining pixels to the non-membrane category. During training, *ilastik* minimises the pixel classification error between this post-threshold category labelling and the ground truth membrane labelling. The final output of *ilastik* is an estimate of the probability that each pixel in the test image is extracellular membrane.

Segmentation and post-processing

To convert the membrane probability map output by *ilastik* into a 2D segmentation, we used a watershed post-processing approach very similar to that used in Andres et al. (2008). The probability map was converted into a set of algorithm-generated objects using the watershed algorithm, using all pixels with a membrane probability below a certain threshold as seeds. Additional post-processing removed all segments below a certain size and all remaining segments were independently subjected to morphological closing. This fills holes and cracks in segments without merging or splitting any segments.

6.7.2 Optimising the ilastik reconstruction

As discussed in section 6.6.1, there are several issues with comparing algorithm performance across studies, and implementations of alternative algorithms are often not available to enable them to be easily compared on the same data set within a study. However, even when studies use available implementations to compare their algorithm against alternative approaches on the same data set, the effort expended in optimising these alternative algorithms for their data set may be unclear. We made an effort to optimise the post-processing that converts the *ilastik* membrane probability map to a 2D segmentation, but we did not optimise any properties of the *ilastik* pixel classifier itself. We summarise the efforts we made to optimise the *ilastik* pipeline below.

Membrane map segmentation

We explored two alternative approaches to converting the membrane probability map output by *ilastik* to a 2D segmentation. The first was a simple *connected components* approach, where the *ilastik* probability map was thresholded to produce a binary map and connected pixels grouped into putative fibre cross-sections. The second was a *watershed* approach (section 6.7.1). For both approaches, the associated threshold was tuned by maximising the *overlap f-measure* achieved against a separate tuning data set (id:4). Overlap was calculated between the algorithm-generated segments and the *all-cell* ground truth polygons (see section 6.7.3). In

both cases, performance was found to be very sensitive to the value of the associated threshold. The optimal watershed approach performed significantly better than the optimal connected components approach (*ilastik B* vs. *ilastik C* in figure 6.11).

Post-processing

Both the connected component and watershed segmentation methods produced small "noise" segments and segments with holes and other morphological defects. To mitigate the effect of these artefacts, we explored the effect of removing all segments below a certain size and the effect of morphologically closing each segment independently. Applying the closing operation to each segment independently ensured that no objects were split or merged during the process and permitted a more aggressive morphological "smoothing" of defects. As with the threshold parameter, minimum object size and number of closing iterations were optimised by maximising the *overlap f-measure* achieved against a separate tuning data set (id:4). Overlap was calculated between the algorithm-generated segments and the *all-cell* ground truth polygons (see section 6.7.3). For both segmentation methods, performance was essentially insensitive to the minimum object size or number of closing iterations.

6.7.3 Making a fair comparison

While comparing the performance of two alternative algorithms on the same data set mitigates many of the issues with inter-study comparisons discussed in 6.6.1, some care must still be taken to make a fair comparison. It is important to ensure that an effort has been made to optimise the alternative algorithm for the comparison data set, and we discuss the steps we took to optimise *ilastik* for our data set in section 6.7.2. We discuss some additional considerations below. The fact that our aim is to only identify a sparse set of *fibre cross-sections*, rather than a produce a dense segmentation of all *cell parts* present in the image, makes a fair comparison especially tricky.

Choice of training task

Ilastik is a pixel classifier. Therefore, when the goal is a sparse segmentation of fibre cross-sections, one approach is to train *ilastik* to perform this task directly. To do this, rather than train *ilastik* to classify pixels as *membrane* or *non-membrane*, we trained it to classify pixels as *fibre* or *other*. The full set of *ilastik* features were used, and all pixels were used for training (data set id:1). Intuitively we would expect *ilastik* to find this task difficult, as many fibre interiors have a similar local appearance to non-fibre regions of the image. This is borne out in our testing, with a fibre interior classifier performing significantly worse than a membrane classifier (*ilastik A* vs. *ilastik B* in figure 6.11). This performance comparison was only made

for the connected components segmentation method.

Choice of ground truth for evaluation

Our algorithm has a different aim than many other algorithms designed to reconstruct neural fibres from electron microscope images. Most 2D algorithms (including *ilastik*) try to identify all *cell parts* in the image, producing a dense labelling of pixels. Our algorithm only attempts to find closed 2D *fibre cross-sections*, producing a sparse labelling of fitted circles. It is not clear what the fairest way to compare these two differing outputs is. It could be argued that the additional objects found by *ilastik* might be more difficult to accurately find than the fibres to which the circle-based algorithm limits itself. On the other hand, it could be argued that finding a subset of cell parts is a more difficult task, combining segmentation and classification. Figure 6.11 shows the performance of various *ilastik* segmentations evaluated against the sparse *fibre-only* ground truth (green) and the dense *all-cell* ground truth (blue). In all cases, *ilastik* performs better on the *all-cell* ground truth. While it could be argued that the *fibre-only* ground truth is the appropriate ground truth for our task, we chose to give *ilastik* the "benefit of the doubt" and evaluate it against the higher scoring *all-cell* ground truth.

An alternative approach would be to evaluate *ilastik* using our measure of *overlap recall*, rather than the *overlap f-measure* that combines both *overlap precision* and *overlap recall*. Recall gives credit for all parts of algorithm-generated objects that overlap the ground truth but does not penalise any parts that do not overlap the ground truth. The recall for the *fibre-only* ground truth was compared for both algorithms and the relative performance of the two algorithms was much the same as that observed when comparing their f-measure scores evaluated against their different ground truths. This suggests that ilastik does not find non-fibres harder to detect than fibres and that, despite the different ground truths used for evaluation, the algorithms can be fairly compared using their respective f-measures.

Region merging

The approach in Andres et al. (2008) is to deliberately produce an over-segmentation of super-pixels from the watershed stage and train another random forest classifier to merge these super-pixels. This functionality is not available in *ilastik*. However, we simulated a "perfect" super-pixel merging algorithm by relaxing the one-to-one matching constraint when scoring the algorithm-generated objects against the ground truth. This has the effect of merging any algorithm-generated objects that maximally overlap with the same ground truth object. However, even after re-tuning the watershed parameters under this relaxed scoring regime, this did not result in a significant improvement in performance. This suggests that the addition of a super-pixel merging stage would not result in a significant improvement in the performance of

ilastik on our data. This may be due to the fact that not all the features used in Andres et al. (2008) are available in *ilastik*. However, it is more likely that this is due to the different tissue stain used in Andres et al. (2008). This extracellular stain does not have the intracellular "clutter" that our classical intracellular stain does. It is possible that a true under-segmentation cannot be produced from the more "cluttered" membrane map that can be generated using the classical stain.

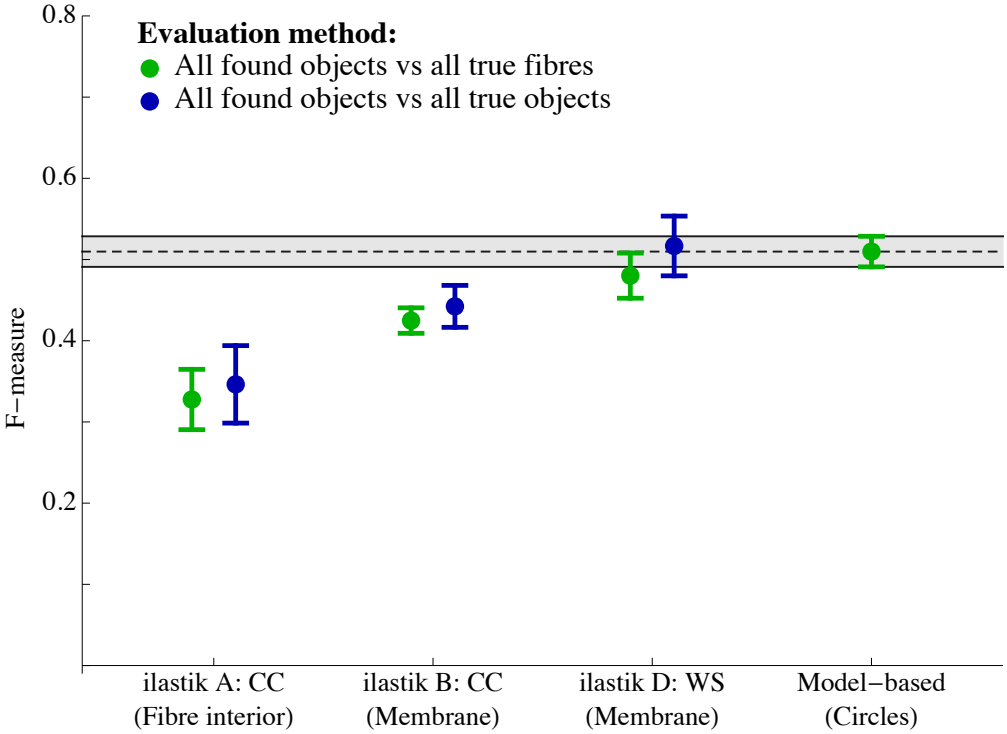


Figure 6.11: Comparing the 2D segmentation performance of our algorithm against that of ilastik, a state of the art pixel-based algorithm. **ilastikA:** ilastik was trained to classify pixels as either *fibre* or *other* and pixels were merged into 2D cross-sections using the *connected-components* algorithm. **ilastikB:** ilastik was trained to classify pixels as either *membrane* or *non-membrane* and pixels were merged into 2D cross-sections using the *connected-components* algorithm. **ilastikD:** ilastik was trained to classify pixels as either *membrane* or *non-membrane* and pixels were merged into 2D cross-sections using the *watershed* algorithm. **Model-based:** our circle-based algorithm. **Green:** algorithm segmentation evaluated against the *fibre-only* ground truth (our desired segmentation). **Blue:** algorithm segmentation evaluated against the *all-cell* ground truth (favouring ilastik).

6.7.4 Ilastik reconstruction accuracy

Figure 6.12 shows the cell segment fibres found by *ilastik* and the corresponding *all-cell* ground truth. Compared to the corresponding visualisation of circle-finding performance in figure 6.8, the most obvious difference is the presence of very large found and ground truth segments. These are associated with Purkinje cell dendrites and glial support cells. Some of these are found well by *ilastik* (e.g. the Purkinje dendrite segments in the upper left and lower right

corners of the image). However, others are not well found, with many of these larger objects merging with several smaller objects. Constraining ourselves to fibre cross-sections, *ilastik* struggles with many of the same large, irregular boutons that our circle-based algorithm does. As *ilastik* relies on reliably identifying extracellular membrane to perform its segmentation, the presence of intracellular membrane signal from vesicles and mitochondria within these boutons is likely to be a significant contributor to this poor performance. However, there are fibres that are well found by one algorithms but not the other.

6.7.5 Comparison of reconstruction accuracy

Across our four test data sets, there is no significant difference between the performance of our circle-based algorithm and the performance of *ilastik*, a state of the art pixel-based classifier (*Model-based* vs. *ilastik D* in figure 6.11). We therefore conclude that the performance of our algorithm is competitive with the state of the art.

6.8 Future work

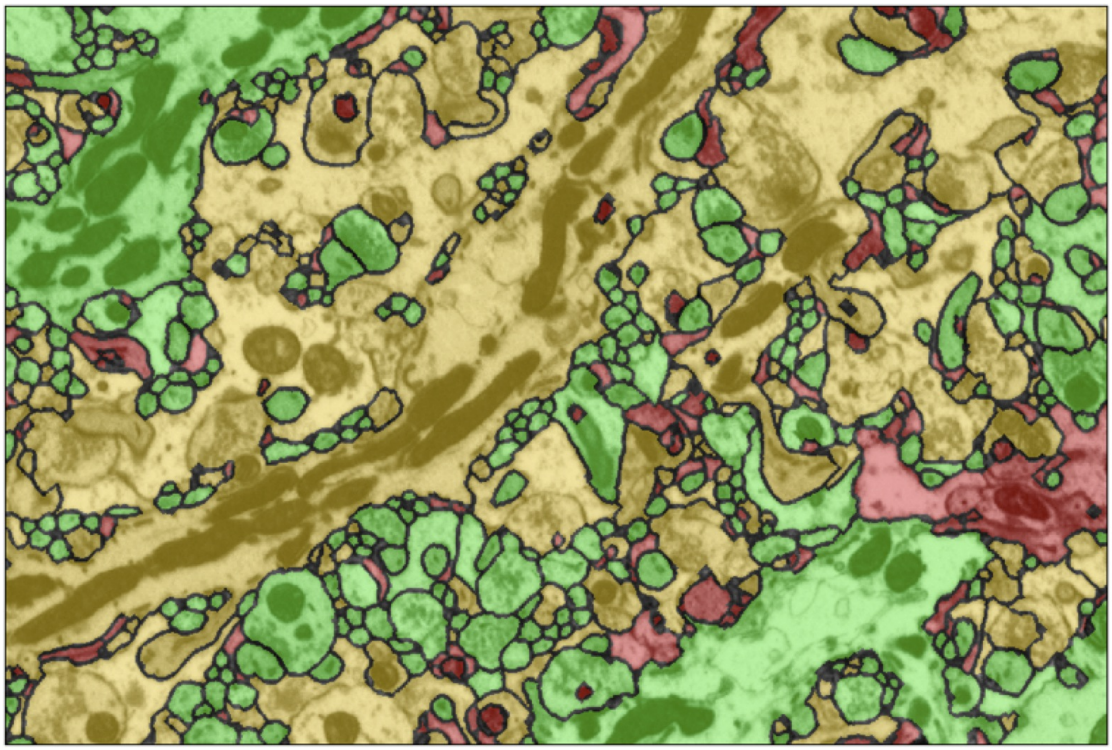
6.8.1 3D features

Extending BIFs to 3D

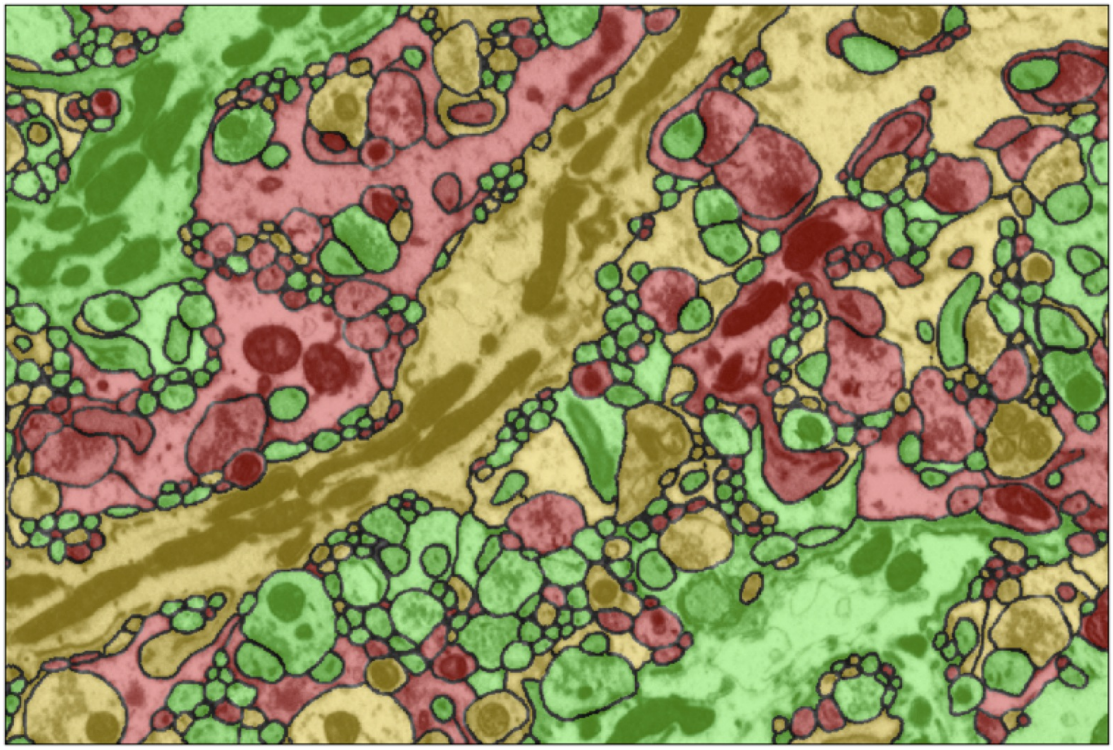
During this work we extended the BIF scheme to use 3D features, which are derived from the second order family of 3D *Derivative of Gaussian* filters. When classifying an image pixel in the current slice, 3D BIFs make use of information from the neighbouring 10 slices on either side. Our 3D BIF scheme has 11 classes. The first two are *flat* and *gradient* classes, while the remaining nine are derived from linear combinations of the 3D Hessian matrix and can be broadly described as *blob*, *pipe*, *sheet* and *saddle* features. In our 3D oBIF scheme, the *gradient* class has 20 quantised orientations and the *pipe*, *sheet* and *saddle* classes have 10. Initial evaluation of 3D BIFs and oBIFs within our 2D fibre finding algorithm resulted in significant improvements in performance when compared to their 2D counterparts. Unfortunately, it was not possible to extend rBIFs to 3D during the course of this work. However, we would expect to achieve a similar increase in performance for 3D rBIFs as we did for BIFs and oBIFs. As 2D rBIF performance exceeded that of 3D oBIFs we continued to use 2D rBIFs for this work.

Using 3D features in ilastik

The performance of *ilastik* also significantly improved when we permitted it to use 3D features. However, as we were unable to extend rBIFs to 3D during the course of this work, our algorithm was limited to considering image evidence from a single slice. To make a fair comparison we therefore limited *ilastik* to 2D features for all evaluation performed in this chapter.



(a) Cell segments found using the best ilastik parameters



(b) Cell segment polygons from manually labelled ground truth

Figure 6.12: **(a)** Cell segments found using the best ilastik parameters. Colour coding indicates *high* (green), *medium* (yellow) and *low* (red) overlap with the ground truth polygons. **(b)** The corresponding manually labelled ground truth polygons colour coded by overlap using the same green/yellow/red scheme to indicate high/medium/low overlap.

6.8.2 Combining algorithms

Both algorithms struggled to successfully identify many of the same large, irregular boutons. However, there were also many fibres that were poorly found by one algorithm but well found by the other. We explored various methods of combining the output of the two algorithms, but were unable to improve performance above that of either algorithm alone. However, further research may be able to discover fruitful methods for combining the output of these two algorithms. We had some success using *ilastik* with 3D features to generate the input to our 2D fibre finding algorithm. However, while this resulted in improved performance, it is likely that this is from the use of 3D features in *ilastik*. If we were to extend rBIFs to 3D we would expect a similar increase in performance without combining the algorithms. However, when we generate circles to use as the input to our 3D algorithm in chapter 7, we use this hybrid approach. This permits us to indirectly make use of 3D features, even though we have not yet managed to extend rBIFs to 3D.

Chapter 7

Reconstructing fibres in three dimensions

In chapter 6 we introduced a novel model-based algorithm for finding circular approximations to neural fibre cross-sections in 2D and demonstrated comparable performance to *ilastik*, a state of the art pixel-based classifier. We now introduce an algorithm that joins the circles found by our 2D algorithm across slices to reconstruct fibres in 3D. We introduce a pair of measures of 3D reconstruction accuracy that capture both the overall proportion of fibre cross-sections successfully reconstructed and the length of successfully reconstructed fibre segments. We then evaluate the performance of our 3D reconstruction algorithm against these measures, discussing parameter sensitivity, the impact of permitting temporary tracking failures and the effect of the finite size of our test volume on our estimates of reconstruction accuracy. We benchmark the performance of our algorithm against results reported by a recently published study that addresses the same problem of reconstructing parallel fibres in mouse cerebellar tissue. Although there are several issues with making a direct comparison with this study, we appear to achieve usefully superior reconstruction accuracy. Finally, we introduce a semi-automated approach, incorporating sparse manual ground truth labelling into our algorithm to improve reconstruction accuracy.

7.1 Algorithm overview

Each 2D image slice is processed as described in chapter 6, generating a fixed set of found fibre circles for each slice, with one key difference. In section 6.8.1 we discussed the improvement in 2D fibre cross-section reconstruction accuracy achieved by the use of 3D features that pool information from multiple adjacent slices. For computational reasons we were unable to extend our 2D annular BIF scheme to directly find local 3D tube segments using 3D rBIFs. However, by training our 2D algorithm using membrane probability maps generated using 3D features in *ilastik*, we are able to indirectly make use of 3D features in our 2D annular BIF scheme. We use this hybrid approach to find the 2D circles used as input to our 3D tube finding algorithm.

To generate 3D found fibre tubes, these 2D circles must be joined across slices. In order to decide whether to join two circles in different slices we must consider two things. Firstly, we must consider how well the circles overlap with one another. We only join two circles if their overlap exceeds a *minimum inter-slice overlap* threshold. Secondly, we must consider how far apart the slices containing the circles are. If we cannot find a suitable match for a circle in the immediately adjacent slice, we can permit ourselves to look for a match in slices that are further away. However, we only join two circles if the separation between their slices does not exceed a *maximum slice separation*. Once we have generated a set of 3D found fibre tubes, we can make a judgement about whether or not we wish to retain them. It is likely that short tubes will not be useful representations of any underlying true fibre and therefore we would be better off discarding them. We therefore enforce a *minimum found fibre length* and discard any found fibre tubes that are shorter than this. The process of joining 2D found fibre circles to form 3D found fibre tubes is described in algorithm 6 and the impact of parameter selection is discussed in section 7.3.

Algorithm 6: Constructing 3D found fibre tubes from 2D found fibre circles

```
Data: Sets of found fibre circles for a series of EM image slices; minimum inter-slice
      overlap (minOverlap); maximum slice separation (maxSep); minimum found
      fibre length (minLength).
Result: Set of 3D found fibre tubes.
assign unique found fibre tube ID to all found circles in slice 1;
for remaining slices do
    for all circles in current slice do
        for previous (maxSep+1) slices do
            calculate overlap between current circle and all circles in previous slice;
            select pairing with maximum overlap;
            if maximum overlap ≥ minOverlap then
                set ID for circle in current slice to ID of circle in previous slice;
                stop processing previous slices;
            end
        end
        if current circle not matched with any previous slices then
            assign new unique found fibre tube ID to current circle
        end
    end
    for all found fibre tubes do
        if number of circles assigned to this tube < minLength then
            delete tube record and all circles associated with the tube ID;
        end
    end
end
```

7.2 Evaluating 3D reconstruction accuracy

In order to evaluate how well our reconstructed 3D fibre tubes represent the 3D ground truth fibres, we consider a pair of measures. The *matched f-measure* indicates the proportion of true and found fibre cross-sections that are *matched* by our algorithm. If there are segments of ground truth fibres where there is no matching segment of found fibre or vice versa, this measure will be reduced. However, the f-measure does not consider how well these matched cross-sections link together to form successfully matched 3D fibre segments. The *matched segment run-length* addresses this, indicating the median length of these successfully matched 3D fibre segments. We discuss these measures further below and relate them to measures reported in other studies.

7.2.1 Matched f-measure

Matched found circles and true polygons

Our threshold for *medium overlap* was set by considering pairs of circles that were *mutually threaded* (section 5.2.1). The concept of *threading* is also used in other studies, with Jurrus et al. (2009a) considering a found cross-section *matched* if it is *threaded* by the corresponding true fibre centreline. We would therefore suggest that our threading-based *medium overlap* threshold is suitable for determining whether found fibre circles should be considered *matched* with true fibre polygons.

Matched f-measure

We evaluated the performance of our algorithm in 2D using the *overlap f-measure*. This continuous measure does not use an overlap threshold for deciding when a found fibre circle and true fibre polygon are *matched*. However, we also considered the proportions of found fibre circles and true fibre polygons having at least *medium overlap* (section 6.5.2). These proportions of *matched* found circles and true polygons can be interpreted as *precision* and *recall* measures. As we did for our continuous *overlap f-measure*, we can take the *harmonic mean* of these two measures to generate a *matched f-measure*. The calculation of this measure is described in algorithm 7. The *matched f-measure* indicates the overall proportion of found fibre and true fibre cross-sections that are matched with each other and is our first measure of 3D reconstruction accuracy. However, it does not consider how well the matched fibre cross-sections are linked together to form 3D matched fibre segments. Lots of small tubes can achieve a *matched f-measure* as high as fewer long tubes if the number of matched cross-sections and overall number of found cross-sections are the same. However, the latter is clearly preferable. In order to distinguish between these cases, we consider the *matched segment run-length*.

7.2.2 Matched segment run-length

Matched segments

Having defined what it means for a found fibre circle and true fibre polygon to be *matched*, we can now define *matched segments*. We define a *matched segment* as a contiguous segment of a 3D found fibre with *all* its constituent cross-sections matched to the same 3D true fibre. However, it can be equivalently defined as a contiguous segment of 3D true fibre with all its constituent cross-sections matched to the same 3D found fibre. The same *matched segments* are identified whether they are determined by considering true fibres or found fibres.

Matched segment run-length

Run-length is the distance over which a 3D fibre is correctly reconstructed, or alternatively the distance between tracking failures. In our case, it is the number of contiguous matched cross-sections comprising a *matched segment*. In order to select the optimal parameters for tube finding, we summarise the run-length distribution by taking the median run-length across all *matched segments*. We use the *median* run-length because the *mean* run-length is ill-defined for our data. This is because some found segments run the full length of our test volume. Note that, in the case where more than half the found segments run the full length of the volume, the median would also be ill-defined. In this case the mean, while still ill-defined, would be more informative. However, this is not the case for our fully-automated reconstruction and we therefore use the median run-length across *matched segments* as our second measure of 3D reconstruction accuracy. The calculation of this measure is described in algorithm 7.

7.2.3 Previously reported measures

Several measures have been used in previous studies to evaluate the accuracy of 3D neurite reconstructions and we discuss some of these below. Figure 7.1 illustrates the calculation of our proposed combined measure and these previously reported measures on a cartoon segmentation example.

Longest matched segment per true fibre

In Jurrus et al. (2009a) only a single *matched segment* is considered for each true fibre, although temporary tracking failures are permitted. Each fibre is seeded either manually or semi-automatically in the first slice, with all found cross-sections in subsequent slices propagated from these initial seeds. Although all found fibres are propagated through all slices, it appears that only the matched segment that includes the initial seed slice is considered when evaluating run-length. It appears that there are additional matched segments for some found fibres, at least when only short tracking failures are permitted. However, these do not appear to contribute

Algorithm 7: Calculating 3D f-measure and run-length.

Data: Set of 3D true fibres; set of 3D found fibres; well-matched overlap threshold (*matchOverlap*).
Result: Matched f-measure; median matched segment run-length.
exclude all *non-interior* true fibres and found fibres (see section 7.2.4);
for *all slices* **do**
  for *all pairings of found fibre circles and ground truth fibre polygons* **do**
    calculate the *overlap* for each polygon-circle pair;
  end
  while *pair with overlap* $\geq$ *matchOverlap remains* **do**
    select pair with highest overlap;
    mark polygon and circle in pair as *matched*;
    increment the *matched count*;
    set overlap for all other pairs involving selected circle and polygon to zero;
  end
end
calculate *precision* by dividing *matched count* by number of found circles;
calculate *recall* by dividing *matched count* by number of ground truth polygons;
calculate *matched f-measure* by taking the *harmonic mean* of *precision* and *recall*;
for *all found fibres* **do**
  collect contiguous sets of *matched* circles into *matched segments*;
  for *all matched segments within a found fibre* **do**
    count number of circles in *matched segment* to calculate *run-length*;
    add *run-length* to list of *all matched run-lengths*;
  end
end
calculate median *matched segment run-length* from list of *all matched run-lengths*;

to the run-length evaluation. Our algorithm generates multiple found fibres matching many of our ground truth fibres. To generate an equivalent run-length measure to Jurrus et al., we would consider only the run length of segments that exist in the first slice of our test volume. However, as our algorithm does not require any seeding of found fibres, there is nothing special about this first slice. Our *matched segment run-length* considers all of the matched segments generated by our algorithm and we would suggest it is a more appropriate measure for our data. Additionally, unlike the Jurrus et al. study, our found fibres can have a range of lengths. Therefore a longer matched segment does not necessarily imply a better overall match. A true fibre matched with a 50-slice found fibre in every slice is a better overall match than a true fibre matched with a 151-slice found fibre for 51 slices. The second found fibre has 100 unmatched slices that are not penalised when considering only the length of the matched segment. Our *matched f-measure* takes such differences into account, penalising unmatched fibre length.

The second issue is that the Jurrus et al. measure only evaluates performance from the perspective of the ground truth fibres. Again, this is reasonable for the Jurrus et al. data, as

there is only ever one found fibre for each true fibre and all fibres are reconstructed through all slices. Therefore the distribution of run-lengths would be identical whether considered from the perspective of the true fibres or the found fibres. However, our algorithm can generate a different number of fibres than ground truth fibres. Again, our combination of *matched segment run-length* and *matched f-measure* provides a symmetric set of measures that takes into account all found fibres and penalises unmatched regions of both found and true fibres.

Longest matched segment per found fibre

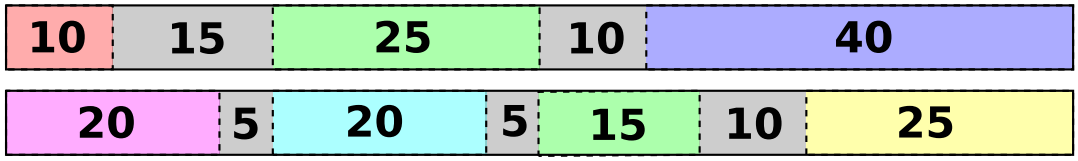
In Jurrus et al. (2013), it appears that the single longest *matched segment* is selected for each found fibre. This is reasonable for their study, as they only attempt to find paths that span all the slices in the volume. However, this measure has the same issues for our data as the corresponding true fibre measure from Jurrus et al. (2009a) discussed above, and our combination of *matched segment run-length* and *matched f-measure* is more appropriate. However, the Jurrus et al. (2013) study addresses the same problem of reconstructing cerebellar parallel fibres as we do. Additionally, they use similar tissue, staining and imaging methods. We therefore benchmark the 3D performance of our algorithm against that reported by this study by attempting to evaluate our results on an equivalent basis. We discuss the issues with making this comparison further in section 7.4.

Number of split and merge errors

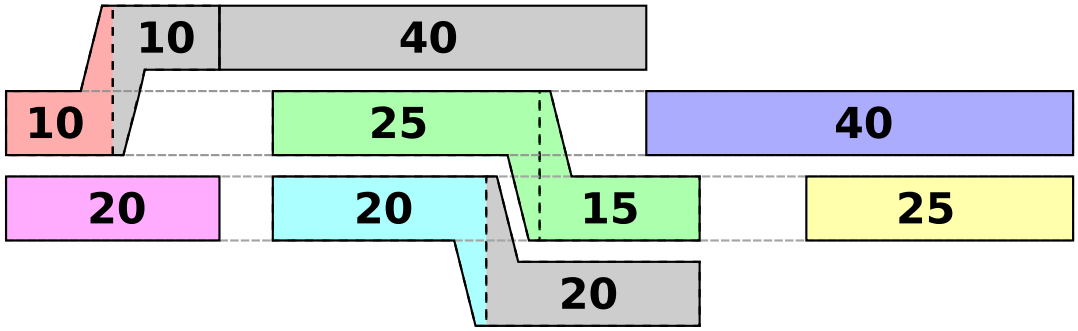
Reported in Turaga et al. (2009, 2010), these measures are calculated for a dense reconstruction where every pixel is labelled. The number of *split errors* is determined by examining the algorithm-assigned labels for all the pixels belonging to each ground truth object. In a perfect reconstruction, all these pixels would have a single algorithm-assigned label. If they do not, then the algorithm has incorrectly split the ground truth object into multiple found objects. The number of *split errors* for a ground truth object is one less than the number of unique algorithm-assigned labels assigned to its pixels. The number of *merge errors* is determined in a similar manner by examining the ground truth labels for all the pixels belonging to each found object. In a perfect reconstruction, all these pixels would have the same label. If they do not, then the algorithm has incorrectly merged multiple ground truth objects into a single found object. The number of *merge errors* for a found object is one less than the number of unique ground truth labels assigned to its pixels.

For dense reconstructions, the densities of split and merge errors are closely related to the run-lengths of matched segments for true and found fibres respectively. However, for sparse reconstructions such as ours, this relationship no longer holds. Most significantly, unmatched segments are not possible in a dense reconstruction. This means that all tracking failures are

accompanied by a split or merge error, which is not the case with sparse reconstructions. Our combination of *matched f-measure* and *matched segment run-length* does not require splits or merges to occur in order to penalise tracking failures and they are therefore a more appropriate pair of measures for sparse reconstructions such as ours.



(a) True fibres



(b) Found fibres

Figure 7.1: Cartoon segmentation example illustrating the calculation of various 3D segmentation measures. In this example there are **(a)** Two true fibres and **(b)** seven found fibres. Coloured segments denote *matched segments*, where a found fibre matches a true fibre for a number of slices. Grey segments denote *unmatched* segments, where there is no match between found and true fibres. Solid lines denote fibre boundaries and dashed lines matched/unmatched segment boundaries within a fibre. The length of each matched/unmatched segment is shown within it. For clarity, each colour in **(a)** corresponds to the same found fibre in **(b)** and the true fibres from **(a)** are shown as dashed great outlines in **(b)**. The various 3D segmentation measures discussed in this section are calculated for this cartoon example as follows. **Our proposed combined measure:** *Matched f-measure:* $155/270 = 0.57$. *Matched segment run-length:* $155/7 = 22.1$. **Longest matched segment run-length for true fibres:** The mean longest matched segment for true fibres is $65/2 = 32.5$. **Longest matched segment run-length for found fibres:** The mean longest matched segment for found fibres is $140/7 = 20$. **Number of split and merge errors:** Using the definitions from Turaga et al. (2009, 2010) and considering all unmatched segments to have the same label: *Splits:* $(4-1)+(5-1) = 7$. *Merges:* $(2-1)+(1-1)+(2-1)+(1-1)+(1-1)+(2-1)+(1-1) = 3$. By dividing the total length of true and found fibres by the number of splits and merges respectively, we can derive corresponding true and found fibre run-lengths of $200/7 = 28.6$ and $225/3 = 75$ respectively.

7.2.4 Mitigating edge effects

When calculating our *matched f-measure* and *matched segment run-length* measures of reconstruction accuracy, we only consider true fibres and found fibre tubes that are *fully contained* within our test volume. By doing this we exclude any tracking failures that are caused by the fibre leaving the volume or touching its edge. We exclude a true fibre as *non-interior* if *any* of

its cross-sections touch the volume edge. We exclude a found fibre as *non-interior* if, for *any* of its circles, the next largest circle is not fully supported within the volume. A circle is fully supported if its centre is at least 1.3× its radius from the edge of the volume.

7.3 Selecting tube finding parameters

There are three parameters associated with the tube-finding process described in algorithm 6. The *minimum inter-slice overlap* determines how much two found fibre circles in different slices must overlap to be joined as part of the same found fibre tube. The *maximum slice separation* determines how many slices can separate two found fibre circles joined as part of the same found fibre tube. The *minimum found fibre length* takes effect after all found fibre tubes have been constructed and removes tubes that do not exceed a minimum length.

We evaluate 3D reconstruction accuracy using a pair of measures. The *matched f-measure* indicates the overall proportion of found fibre and true fibre cross-sections that are matched with each other, while the median *matched segment run-length* is the average length of fibre segments comprised of contiguous matched cross-sections. Unfortunately it is not possible to simultaneously maximise both these measures by selecting appropriate values for our three parameters. We discuss the sensitivity of both measures to each of our tube finding parameters and the trade-off that may be made between the two measures.

7.3.1 Initial parameter sensitivity exploration

In order to understand the impact of each parameter on tube finding performance, we will first select the optimum *minimum inter-slice overlap* and *maximum slice separation* for the case where we do not enforce a *minimum found fibre length*. We then examine the effect of enforcing a *minimum found fibre length*, discovering that it is this parameter that makes it impossible to simultaneously optimise both *matched f-measure* and median *matched segment run-length*.

No minimum found fibre length

When no *minimum found fibre length* is enforced, it is possible to set the *minimum inter-slice overlap* and *maximum slice separation* to maximise the median *matched segment run-length* with little impact on the *matched f-measure*. Figure 7.2 shows a clear maximum run-length of 18 slices when the *minimum inter-slice overlap* is set to 0.3 (7.2a) and the *maximum slice separation* is set to 20 (7.2c). In contrast, the f-measure exhibits very little sensitivity to these two parameters.

Enforcing a minimum found fibre length

However, if we consider the effect of varying the *minimum found fibre length* on our pair of accuracy measures, the effect is less straightforward (figure 7.3). Keeping the other parameters

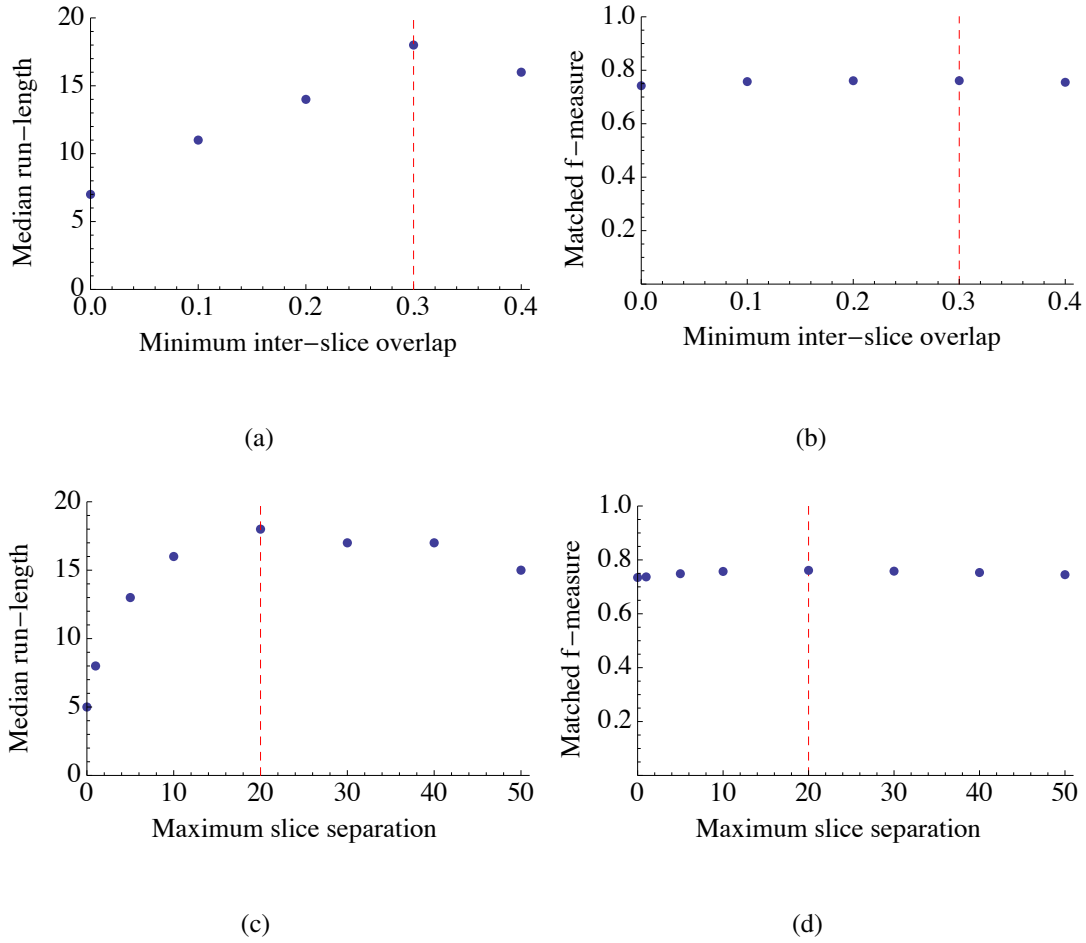
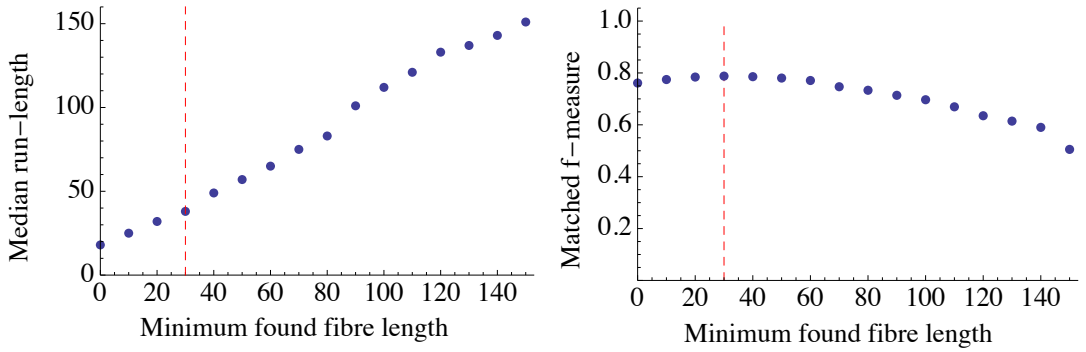


Figure 7.2: Tube finding parameter sensitivity with no minimum fibre length. **(a)** and **(c)** *matched segment run-length* has a clear maximum of 18 slices with a *minimum inter-slice overlap* of 0.3 and a *maximum slice separation* of 20. **(b)** and **(d)** *matched f-measure* is relatively insensitive to variations in either parameter. Dashed red lines indicate optimum parameter values.

fixed at the values shown in figure 7.2, increasing the minimum found fibre length from zero results in a monotonic increase in median *matched segment run-length* (7.3a). However, the *matched f-measure* begins to noticeably fall once the minimum fibre length is increased beyond ∼40 slices (7.3b). Interestingly, the f-measure actually rises slightly at first. This is because very short found fibres tend to be poorly matched with ground truth fibres. Therefore removing these fibres eliminates more *unmatched* cross-sections than *matched* cross-sections. As longer found fibres are removed, the number of *matched* cross-sections eliminated increases, resulting in a reduction in f-measure.

Increasing the minimum fibre length from 0 to 30 slices results in the median run-length almost tripling from 18 to 49 slices, while the f-measure increases slightly from 0.76 to 0.79. This is clearly an unambiguous improvement. Given that the rapid rise in median run-length beyond this point is accompanied by a relatively slow fall in f-measure, it could also be argued that it would be reasonable to trade-off a small decrease in the latter for a large increase in the former. For example, increasing the minimum fibre length to 70 slices would increase the median run-length to 75 slices at the cost of reducing the f-measure to 0.75. This is within 5% of the maximum achievable and only marginally below that achieved with no *minimum found fibre length*. We explore this trade-off between *matched segment run-length* and *matched f-measure* further in section 7.3.2.



(a) (b)

Figure 7.3: Effect of enforcing a minimum fibre length. **(a)** *matched segment run-length* increases monotonically as the *minimum found fibre length* is increased. **(b)** *matched f-measure* initially increases slightly before gradually decreasing as *minimum found fibre length* is increased

7.3.2 Balancing matched segment run-length and matched f-measure

In section 7.3.1 we fixed the *minimum inter-slice overlap* and *maximum slice separation* at their optimum values when no *minimum found fibre length* was enforced. We found that both

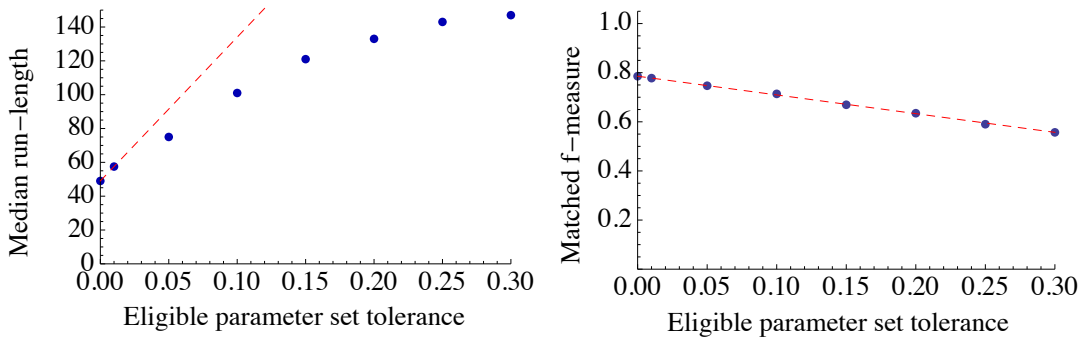
the median *matched segment run-length* and the *matched f-measure* initially improved as the minimum found fibre length was increased. After this point the run-length continued to improve while the f-measure began to decline. However, as this decline was relatively slow it would be possible to increase the achieved run-length further at a small cost to the achieved f-measure. Here we more fully explore the trade-off between median *matched segment run-length* and *matched f-measure* while permitting all three tube finding parameters to vary.

Limiting parameter selection to achieve a minimum f-measure

We calculated the run-length and f-measure achieved for a wide range of combinations of *minimum inter-slice overlap*, *maximum slice separation* and *minimum found fibre length*. We determined the maximum f-measure achieved across all these parameter sets and then used this as a reference point for selecting *eligible parameter sets* with varying tolerance. For example, a tolerance of 0% selects only the parameter sets that achieve the maximum f-measure of 0.79, while a tolerance of 5% selects all parameter sets that achieve an f-measure of 0.75 or greater. From all the *eligible parameter sets* selected for a given tolerance, we then selected the single parameter set that maximised median run-length. Figure 7.4 illustrates the trade-off between *matched segment run-length* and *matched f-measure* as the tolerance used to select *eligible parameter sets* is varied from 0-30%. It can be seen that the median run-length of matched segments can be increased at the cost of a reduction in the overall proportion of segments matched. The f-measure falls linearly with increasing tolerance, while the median run-length rises sub-linearly with increasing tolerance. Depending on the relative value assigned to these increases in run-length and decreases in f-measure, it may be reasonable to make a trade-off between the two by permitting a non-zero tolerance for selecting *eligible parameter sets*. However, for this work we have chosen to be conservative and require our selected parameter set to have the maximum achievable f-measure of 0.79. Therefore our maximum achievable run-length is 49 slices, achieved at a *minimum inter-slice overlap* of 0.3, a *maximum slice separation* of 20 and a *minimum found fibre length* of 40.

Visualising reconstruction accuracy for found and true fibres

The effect of enforcing a *minimum found fibre length* can be further understood by visualising how well individual found and true fibres are matched. This is illustrated in figure 7.5 for found fibres and figure 7.6 for true fibres. Each column represents a fibre and each row a slice. Black indicates that the fibre does not exist at that slice, red indicates that the fibre exists but has not been matched and green indicates that the fibre has been matched. Blue lines mark points where the identity of the matched fibre changes with no intervening unmatched slices. Therefore sections of green separated by either blue lines or sections of red are separate *matched*



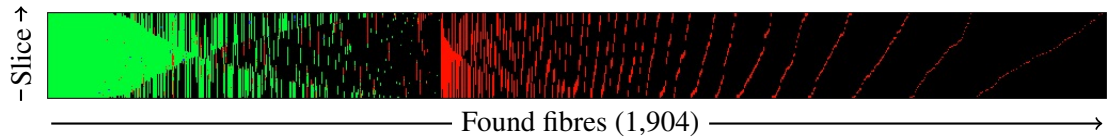
(a) (b)

Figure 7.4: Run-length vs. matched f-measure trade-off. As the tolerance used to set the f-measure threshold for the selection of *eligible parameter sets* is increased, a trade-off occurs between run-length and f-measure. **(a)** *matched segment run-length* increases sub-linearly with increasing tolerance. **(b)** *matched f-measure* decreases linearly with increased tolerance. Dashed red lines are a linear extrapolation from the first two points on each plot.

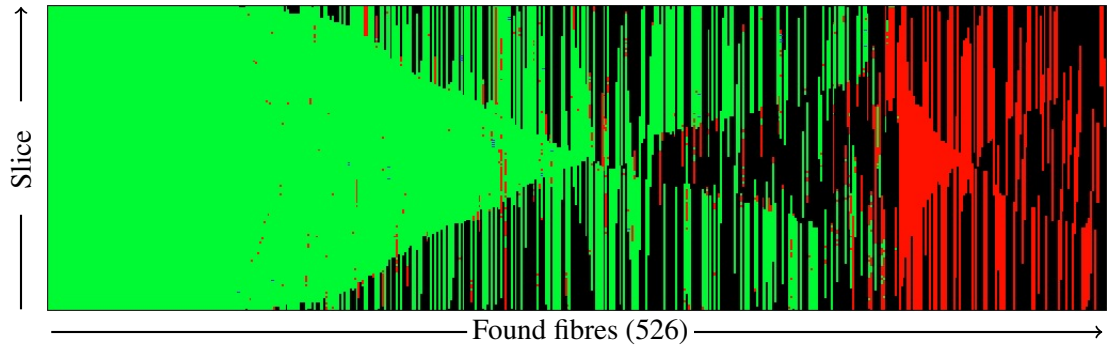
segments. Note that the left-to-right ordering of fibres can differ between plots as fibres are ordered by number of matched cross-sections for clarity of presentation. Also, figure 7.5a is shown at ∼28% of the scale of figures 7.5b and 7.5c.

The effect of enforcing a minimum found fibre length

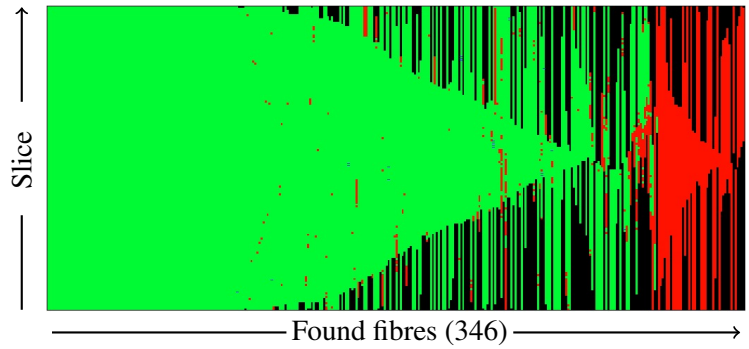
The effect of enforcing a minimum found fibre length is most apparent when considering the range of fibres found with various minimum found fibre lengths (figure 7.5). With no minimum found fibre length (7.5a) 1,904 fibres are found, with many of them being very short and unmatched to any of the 377 true fibres. These comprise the entire right hand half of the plot. By selecting the set of parameters that maximise the *matched f-measure* (7.5b), most of these short unmatched found fibres are discarded. This change of parameters also results in a significant increase in median *matched segment run-length* from 18 to 49 slices. However, this is not a result of discarding the short unmatched found fibres, as these are not considered in the run-length calculation. The cause of the increase in run-length can be seen by examining how well true fibres are matched (figure 7.6). Comparing the matched segments with no minimum found fibre length (7.6a) to those with maximised f-measure (7.6b), three main effects can be observed. Firstly, there are fewer blue lines. This corresponds to fewer cases where the found fibre matched to a true fibre changes. Many of the blue lines eradicated by this change in parameters are pairs of lines that bracket short stretches where the long-term matched found fibre is temporarily displaced by another shorter fibre that is a better match in a few slices. Removing these short fibre segments results in the long-term matched found fibre being the best match at these points, and results in a single longer matched segment replacing multiple smaller ones.



(a) No minimum found fibre length enforced



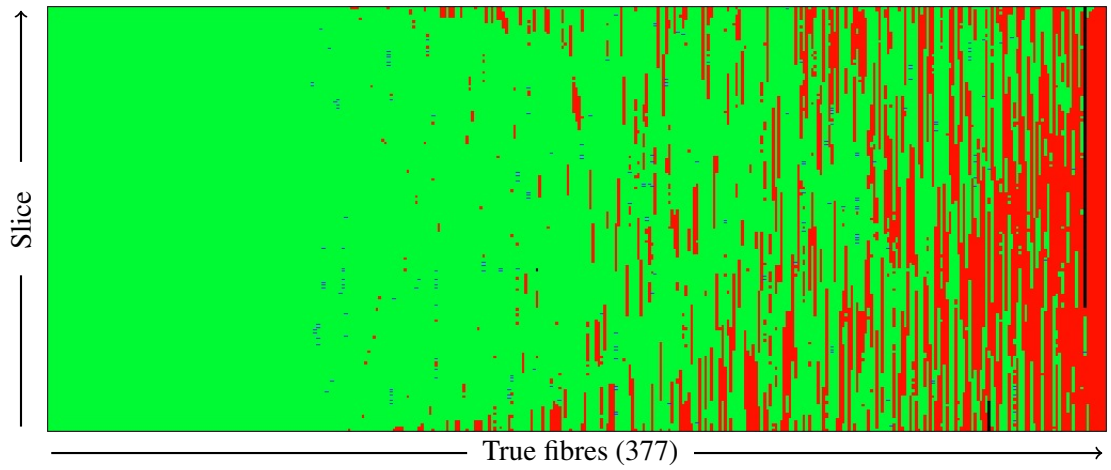
(b) Minimum found fibre length selected to maximise *matched f-measure*



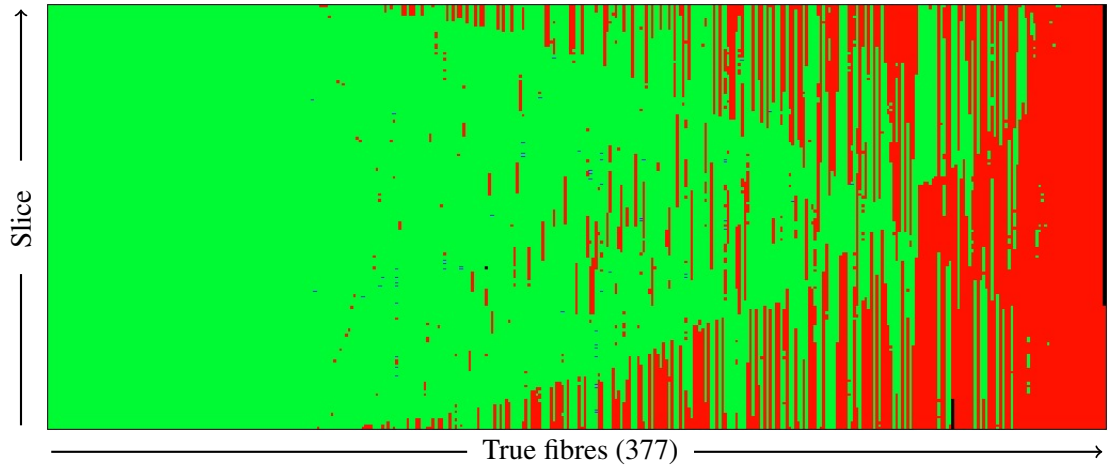
(c) Minimum length selected to maximise *run-length* while keeping *f-measure* within 5% of maximum

Figure 7.5: Matched segments for found fibres. Black indicates that the fibre does not exist at that slice. Red indicates that the fibre exists but has not been matched to a segment of true fibre. Green indicates that the fibre has been matched to a segment of true fibre. Blue lines mark points where the identity of the matched true fibre changes with no intervening unmatched slices. Therefore sections of green separated by either blue lines or sections of red are separate *matched segments*. Note that the left-to-right ordering of fibres can differ between plots as fibres are ordered by number of matched cross-sections for clarity of presentation. Also, figure 7.5a is shown at ∼28% of the scale of figures 7.5b and 7.5c.

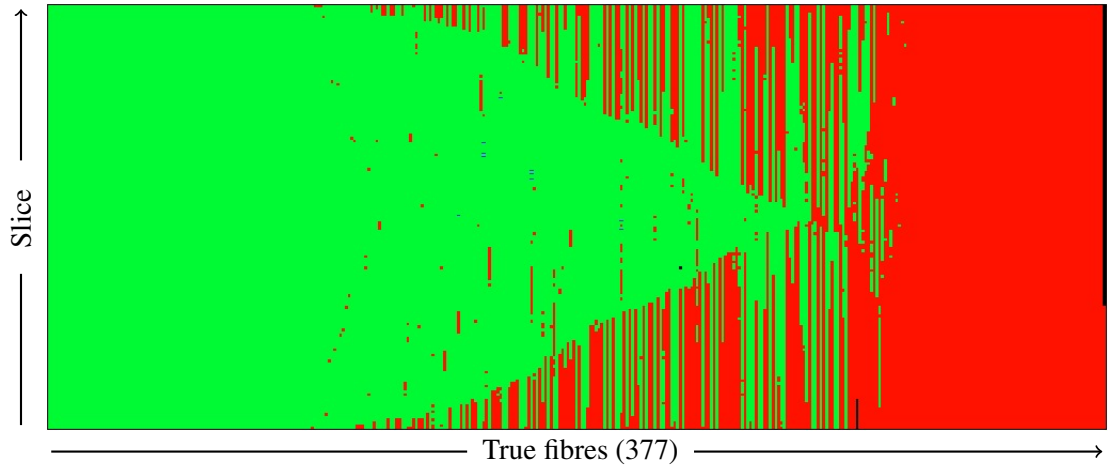
Secondly, there is more red at the right side of the plot. This corresponds to more fibres that are completely unmatched due to the discarding of short fibres that contributed correspondingly short matched segments. Thirdly, the distribution of green in the plot becomes more triangular in the centre. This corresponds to some matched segments disappearing where true fibres were previously matched to multiple found fibres, with the shorter fibres and their correspond-



(a) No minimum found fibre length enforced



(b) Minimum found fibre length selected to maximise *matched f-measure*



(c) Minimum length selected to maximise *run-length* while keeping *f-measure* within 5% of maximum

Figure 7.6: Matched segments for true fibres. Black indicates that the fibre does not exist at that slice. Red indicates that the fibre exists but has not been matched by a segment of found fibre. Green indicates that the fibre has been matched by a segment of found fibre. Blue lines mark points where the identity of the matched found fibre changes with no intervening unmatched slices. Therefore sections of green separated by either blue lines or sections of red are separate *matched segments*. Note that the left-to-right ordering of fibres can differ between plots as fibres are ordered by number of matched cross-sections for clarity of presentation.

ing shorter matched segments being discarded. All these effects act to increase the *matched segment run-length* by preferentially dropping fibres with shorter matched segments, while the latter two effects also act to decrease the *matched f-measure*.

The initial increase in *matched f-measure* observed when enforcing a *minimum found fibre length* is due to the elimination of a large number of *unmatched* cross-sections associated with the discarded short fibres. This is apparent in a reduction in the overall amount of red between figures 7.5a and 7.5b. However, this is partially offset by the elimination of the *matched* cross-sections associated with the discarded short fibres. This is apparent in a reduction in the overall amount of green between figures 7.6a and 7.6b. Note that there is an equivalent reduction in the overall amount of green between figures 7.5a and 7.5b, but it is difficult to see due to the different scales used for the figures. When increasing the *minimum found fibre length* further in order to increase the *matched segment run-length*, the overall number of matched and unmatched cross-sections are both further reduced (compare the overall amounts of green and red between figures 7.5a and 7.5b or between figures 7.6a and 7.6b). However, the number of matched cross-sections eliminated increases faster than the number of unmatched cross-sections, resulting in a reduction in *matched f-measure* as the *minimum found fibre length* is increased further.

An argument could be made that it would be reasonable to increase the run-length further at the cost of a small reduction in f-measure. For example, by increasing the *minimum found fibre length* from 40 to 70 slices, the median *matched segment run-length* is increased from 49 to 75 slices, at the cost of an $\sim$5% reduction in *matched f-measure* from 0.79 to 0.75. However, for this work we have chosen to be conservative and require our selected parameter set to have the maximum achievable f-measure of 0.79. Therefore our maximum achievable median run-length is 49 slices. Figure 7.7 illustrates the sensitivity of both median run-length and f-measure as each of our three tube finding parameters are varied around our chosen *maximum f-measure* values.

7.3.3 Visualising individual reconstructed fibres

Figure 7.8 illustrates the performance of our algorithm using our most conservative *maximal f-measure* parameters. It shows 12 example true fibres (blue), along with all found fibres that are matched with them (green and cyan). True fibres are ordered by total number of matched segments, then by length of longest matched segment. Therefore lower numbered fibres are generally better matched by their corresponding found fibres. This is the same ordering used in figure 7.6b, with *true fibre n* in figure 7.8 being the n^{th} column in figure 7.6b. The 12 examples are evenly spaced 32 fibres apart and therefore are a reasonably representative illustration of

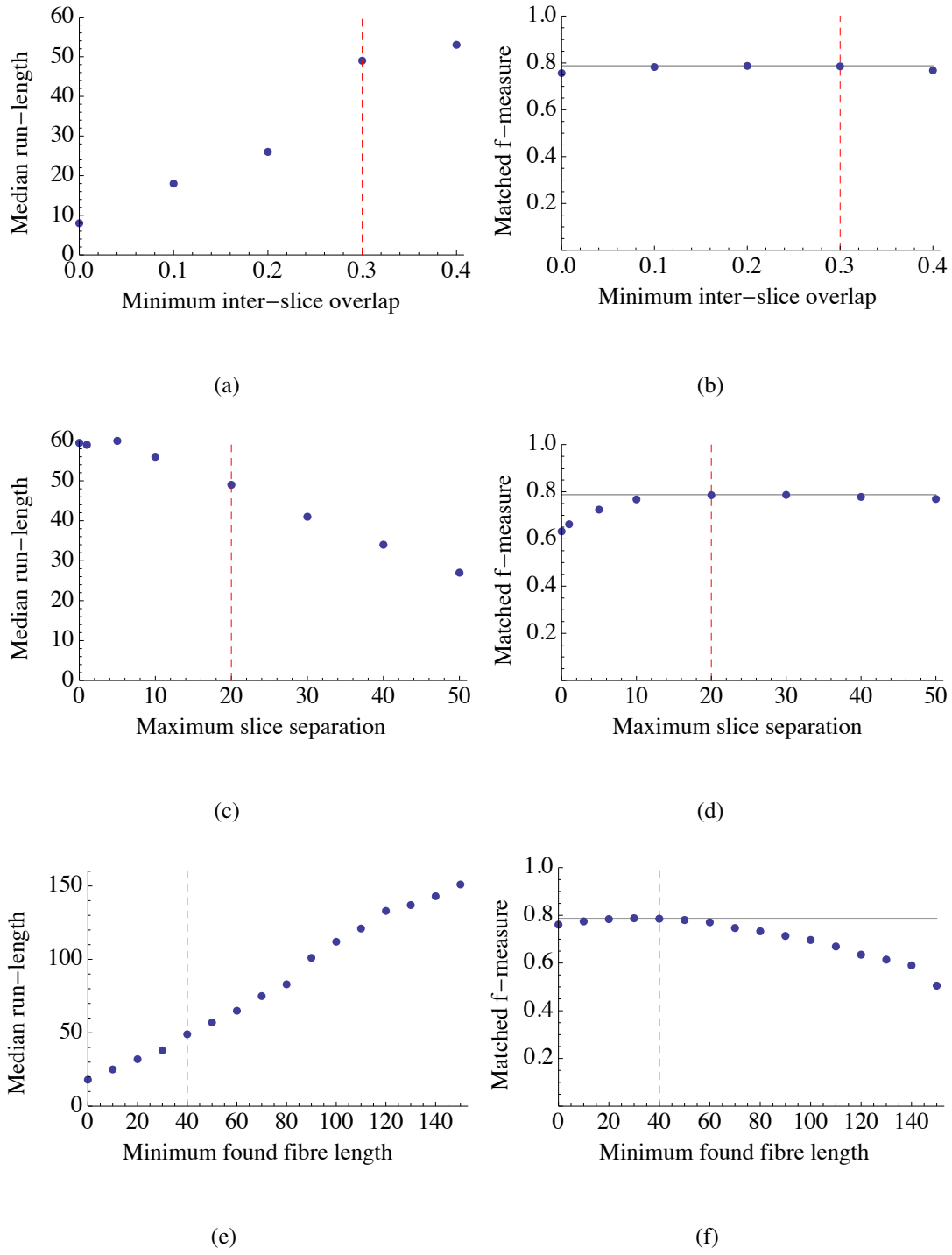


Figure 7.7: Tube finding parameter sensitivity with an enforced minimum found fibre length. When the *minimum found fibre length* is permitted to vary, there is no set of parameters that simultaneously maximise *matched segment run-length* and *matched f-measure*. However, as shown in figure 7.4, a trade-off can be made between these two measures, depending on how much we permit the f-measure to fall below its maximum achievable value. We have been conservative and selected the parameters that maximise the f-measure and these plots show the effect of varying each of the three tube finding parameters while keeping the other two at their *maximum f-measure* values. **Left:** the effect of varying each parameter on the median *matched segment run-length*. **Right:** the effect of varying each parameter on the *matched f-measure*. Red dashed lines indicate the *maximum f-measure* parameter values.

the full range of 3D fibre finding performance. Qualitatively, good one-to-one matches between true and found fibres appear to be made up to fibre number 129. However, overall only 94 of the 377 true fibres are *fully tracked* through the test volume (i.e. are matched by the same found fibre in all 151 slices). Some of the low numbered fibres will be matched to a single found fibre but suffer from a temporary tracking failure, where a few slices do not achieve sufficient overlap to be considered *matched*. These will appear to be very good matches in figure 7.8 but will count as two shorter *matched segments* when run-lengths are calculated. We discuss the impact of permitting temporary tracking failures in section 7.3.5. As fibres are ordered by total number of matched cross-sections, some split fibres such as true fibre 158 are also likely to be present among the low number fibres. However, overall the visualisation of individual fibres provides a view of fibre finding accuracy that is consistent with the overview presented in figure 7.6b.

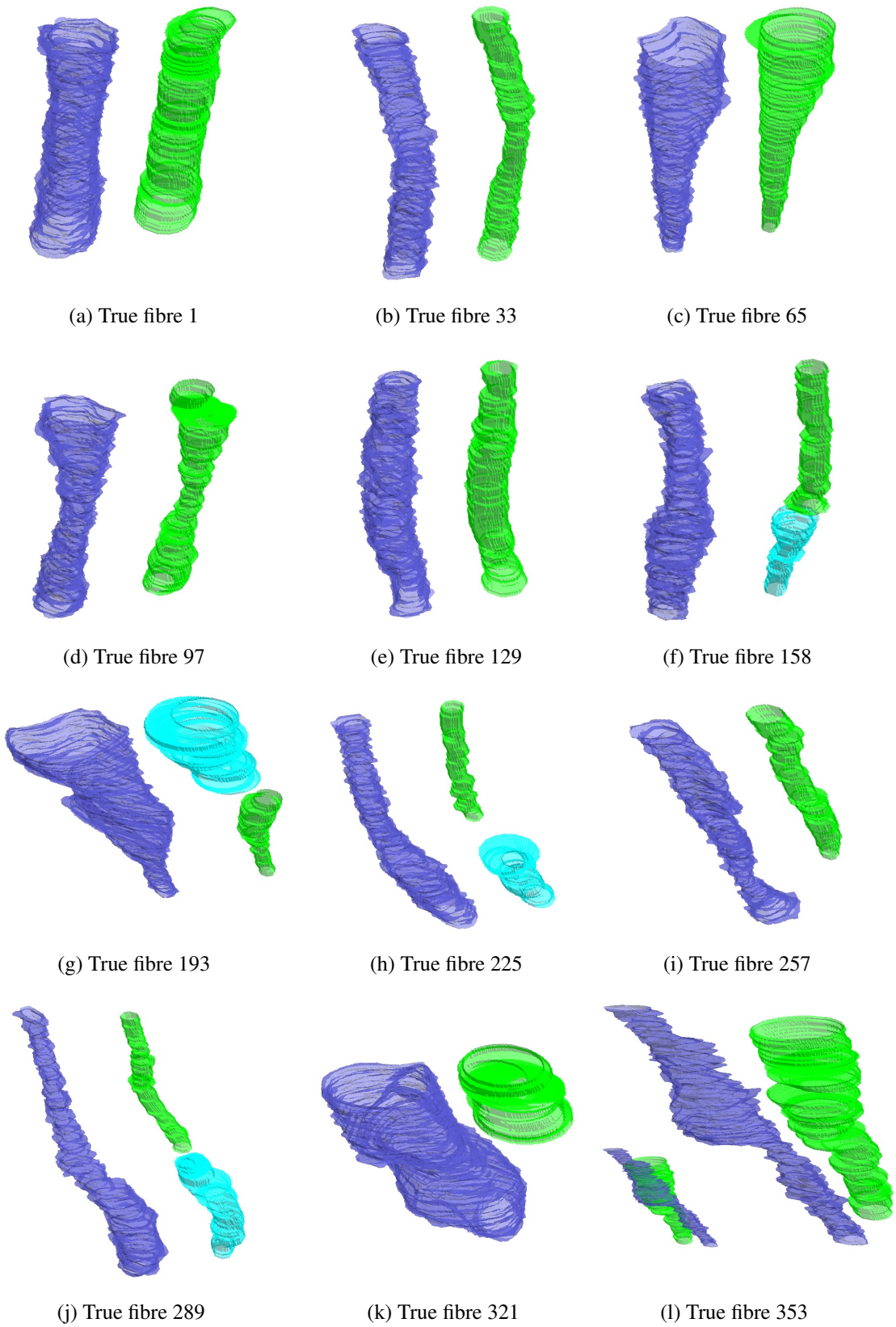
7.3.4 Summarising 3D reconstruction accuracy

Previous studies have reported fibre tracking performance by plotting run-length *survival functions* (Jurrus et al., 2009a, 2013). These plots show the proportion of objects exceeding each observed run-length, and the Jurrus et al. studies plot survival functions for true and found fibres respectively. We plot similar survival functions for *matched segments* in figure 7.9. As discussed in section 7.3.2, it is possible to increase the median *matched segment run-length* at the expense of a reduction in the overall proportion of matched cross-sections (*matched f-measure*). Figure 7.9 shows survival functions for the three *minimum found fibre lengths* examined, summarising the run-length information presented in figures 7.5 and 7.6.

7.3.5 Permitting temporary tracking failures

Jurrus et al. (2009a) permit temporary tracking failures, generating survival functions using two criteria for tracking recovery. Their *Metric A* considers a fibre to be continuously tracked if a tracking failure is recovered by the end of the fibre. Their *Metric B* considers a fibre to be continuously tracked if a tracking failure is recovered within 10 slices (500 nm). We explored the effect of applying similar criteria to permit recovery from temporary tracking failures for our data, using our *maximum f-measure* parameters. Figure 7.10 shows the *matched segment run-length* survival functions achieved for our *maximum f-measure* parameters by permitting temporary tracking failures that recover after 1, 5, 10 slices (9.3, 46.5 or 93 nm). It also shows the survival functions if no tracking failures are permitted and if temporary failures that recover by the end of the fibre are permitted.

Permitting tracking failures as short as a single slice results in a significant improvement in



(a) True fibre 1 (b) True fibre 33 (c) True fibre 65

(d) True fibre 97 (e) True fibre 129 (f) True fibre 158

(g) True fibre 193 (h) True fibre 225 (i) True fibre 257

(j) True fibre 289 (k) True fibre 321 (l) True fibre 353

Figure 7.8: Example true fibres (**blue**) and their corresponding fully-automated found fibres (**green** and **cyan**). True and found fibres are separated for clarity. True fibres are ordered as in figure 7.6b, with *true fibre n* being the n^{th} column in figure 7.6b. Longest found fibres are in green, with additional fibres in cyan. The quality of the overall match between true and found fibres falls from (a) to (l), with examples evenly spaced 32 fibres apart. The exception is fibre 158 which was chosen to more accurately reflect the proportions of split fibres. The fibres in (l) may look like a good match displayed side by side, but they are not well aligned (see inset).

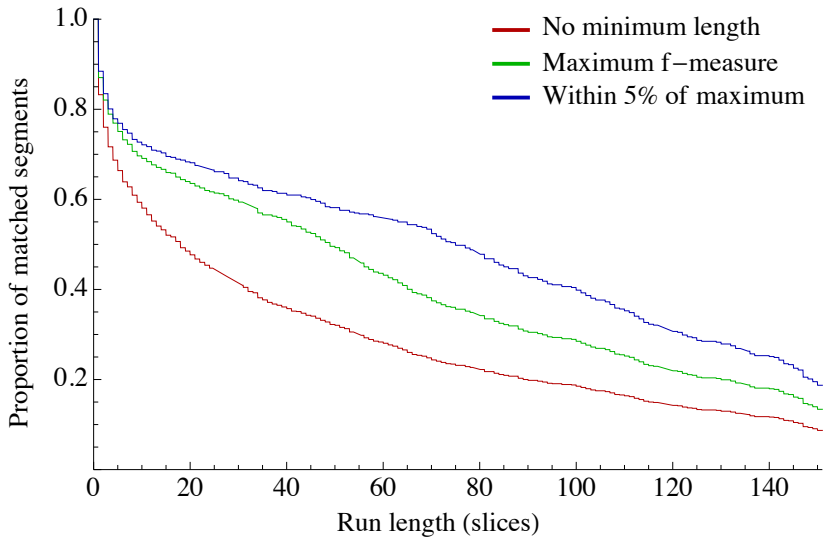


Figure 7.9: Run-length distribution for a selection of parameter sets. Enforcing a *minimum found fibre length* while maximising the f-measure (green) results in a significant improvement in the run-length distribution compared to having no *minimum found fibre length* (red). Increasing the *minimum found fibre length* further, while maintaining the f-measure within 5% of its maximum results in a further significant improvement to the run-length distribution (blue).

median *matched segment run-length* from 49 to 71 slices, while permitting temporary tracking failures for 5 and 10 slices results in further increases in median run-length to 83 and 88 slices respectively. Permitting all temporary tracking failures that recover by the end of the fibre only results in a small further increase in median run-length to 90.5 slices. The *matched f-measure* remains at 0.79 in all cases, as permitting temporary tracking failures does not affect the number of well-matched fibre cross-sections. The decision whether to permit temporary tracking failures only affects the run-lengths reported. The actual found fibres remain the same, whether temporary tracking failure is permitted or not. We would argue that permitting temporary tracking failures of up to 93 nm (10 slices) is reasonable, given that the ultimate goal is to track parallel fibres over millimetres. The key question is whether these temporary tracking failures result in reconstructed models that have significantly different properties from a neural modelling perspective. The found fibre is still connected over the extent of these tracking failures, so the within-fibre connectivity is not changed by permitting them. The only risk to recovering accurate connectivity is that of failing to identify a potential pre-synaptic bouton. As the 93 nm tracking failures we are proposing are 5-15$\times$ shorter than the typical length of a pre-synaptic bouton it is unlikely that a tracking failure within a bouton will prevent it being identified. Such short temporary failures are therefore unlikely to cause any changes to the connectivity of the reconstructed network. While the mis-estimation of fibre position and radius at tracking failures may result in some differences in the reported physical or electrotonic length of fibres it is likely

that these differences will be small. Compartments in a multi-compartment model are likely to be significantly longer than 93 nm. Therefore any errors in cross-section position and radius introduced by temporary tracking failures will likely be averaged out when the poorly matched cross-sections comprising a short tracking failure are combined with the longer successfully matched cross-sections from either side.

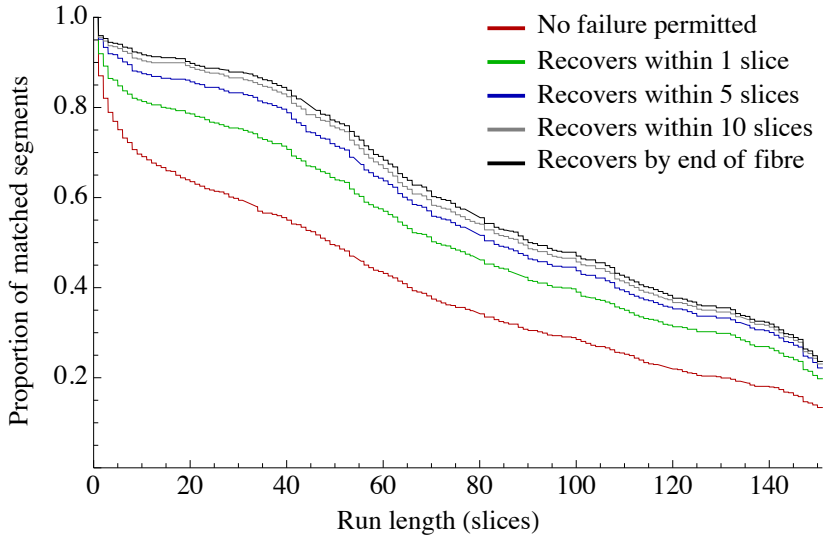


Figure 7.10: Effect of permitting temporary tracking failures on the *matched segment run-length* survival curve for our *maximum f-measure* parameters. Permitting tracking failures as short as a single slice results in a significantly improved survival function. Smaller but still noticeable improvements are seen when the maximum permitted tracking failure is increased further to 5 and 10 slices. However, the additional improvement seen by permitting all tracking failures that recover by the end of the fibre is relatively small.

7.3.6 Accounting for censored data

Data is *right-censored* when its value is known to be *above* a certain value, but its true value is unknown. Similarly, data is *left-censored* if its value is known to be *below* a certain value but its true value is not known. *Left censoring* is very rare. As we are measuring the run-length of *matched segments* within a finite volume, our run-length data suffers from *right-censoring*. *Right-censoring* of our run-length data occurs when a *matched segment* leaves the volume at one of the edges. When this occurs, the measured length of the segment is actually a *lower bound* on its true length. As a result, the run-length survival function will underestimate the proportion of segments exceeding each run-length. For our data, a *matched segment* is right-censored if it contains a *matched cross-section* in either the first or the last slice of our test volume. For our *maximum f-measure* parameters, $\sim$50% of matched segments are right-censored when no temporary tracking failure is permitted. When temporary tracking failures of up to 10 slices are permitted, the proportion of right-censored matched segments rises to

~70%. Fortunately, right-censored data can be accounted for in a relatively straightforward manner using the Kaplan-Meier method (Kaplan and Meier, 1958). Figure 7.11 shows the effect of accounting for censoring in this manner for our data. It can be seen that the effect of right-censoring on our run-length estimates is significant. For our *maximum f-measure* parameters with no temporary tracking failure permitted, the median run-length rises from 49 to 81 slices. While the estimate of median run-length from the Kaplan-Meier survival function is considered unbiased (Zhong and Hess, 2009), if the data is heavily censored the median run-length may be greater than the number of slices in the test volume. This is the case when temporary tracking failures of up to 10 slices are permitted for our *maximum f-measure* parameters. In this case we can only say that the median *matched segment run-length* is greater than 151 slices. In order to generate a Kaplan-Meier estimate for the true median run-length we would need to evaluate our algorithm on a larger test volume.

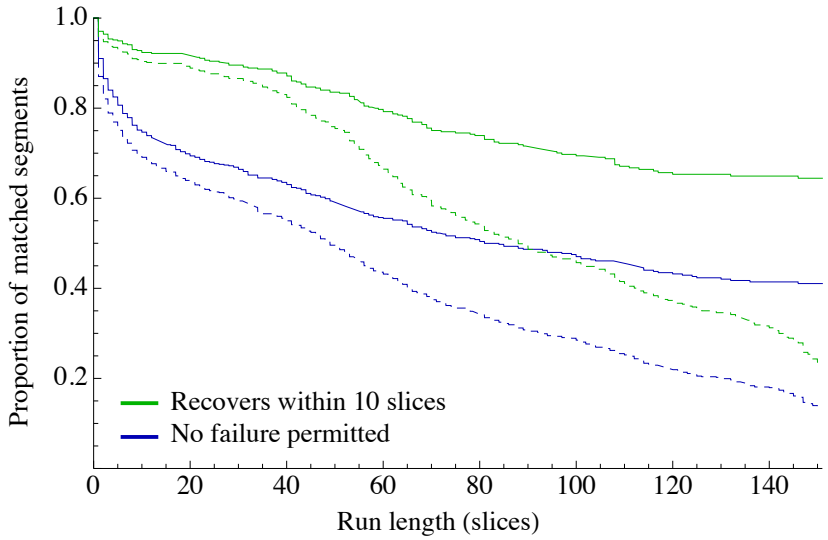


Figure 7.11: Effect of accounting for the right-censoring of *matched segment run-lengths* using the Kaplan-Meier method. Survival functions are shown for our *maximum f-measure* parameters with no temporary tracking failure permitted (**blue**) and tracking failures of up to 10 slices permitted (**green**). Dashed lines are the original survival functions and solid lines are the Kaplan-Meier survival functions. The effect of censoring on our run-length estimates is large.

To attempt to estimate the reliability of the Kaplan-Meier method for our data, we applied it to truncated subsets of our data consisting of the first 139, 101 and 51 slices of our 151 slice volume. For the 139 slice subset, the Kaplan-Meier median run-length estimate increased to 125 slices (compared to 81 for the full 151 slice volume). For subsets containing less than 139 slices the Kaplan-Meier median run-length estimate becomes undefined, as more than 50% of the matched segments span all the slices in the sub-volume. However, examining the survival functions for these Kaplan-Meier adjusted run-lengths (figure 7.12), it appears that the Kaplan-

Meier adjustment is likely to increasingly over-estimate the true run-lengths as the proportion of censored segments increases. It is therefore likely that the true run-length survival function for our full 151 slice data set lies somewhere between the unadjusted (black) and Kaplan-Meier adjusted (grey) curves shown in figure 7.12). In order to avoid overestimating the performance of our algorithm when benchmarking (section 7.4), we use the unadjusted run-length data and do not apply the Kaplan-Meier adjustment.

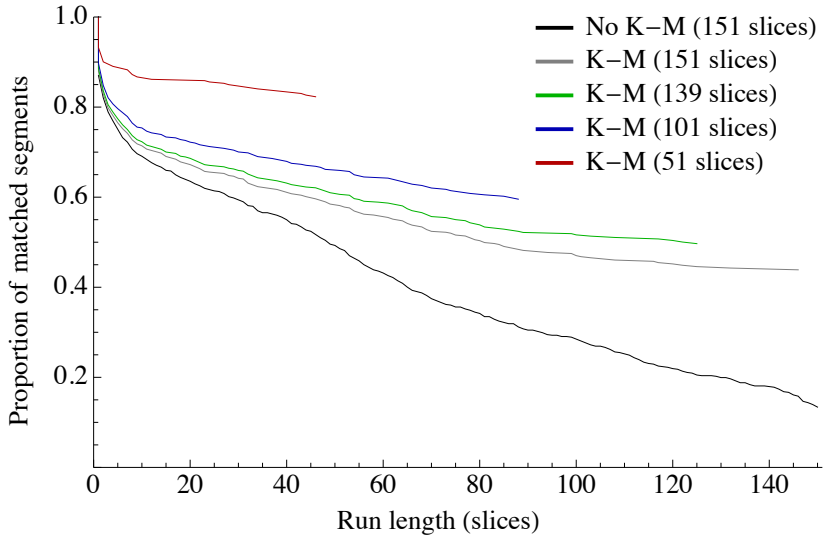


Figure 7.12: Validating the Kaplan-Meier estimation method on our data by truncating our reconstruction in z. **Black:** Run-length survival function for full 151 slice volume with no Kaplan-Meier adjustment. **Grey:** Kaplan-Meier adjusted run-length survival function for full 151 slice volume. **Other colours:** Kaplan-Meier adjusted run-length survival functions for sub-volumes truncated to the first 139 (**green**), 101 (**blue**) and 51 (**red**) slices. The Kaplan-Meier median run-length estimates for the 151 and 139 slice volumes are 81 and 125 slices respectively. Due to the increasing proportion of segments that span the entire sub-volume as the number of slices is reduced, the Kaplan-Meier median run-length estimate becomes undefined for sub-volumes containing fewer than 139 slices. For the 101 and 51 slice sub-volumes, we can only say that the Kaplan-Meier median run-length estimate is greater than 88 and 46 slices respectively. However, examining the full family of survival functions, it appears that the Kaplan-Meier adjustment is likely to increasingly overestimate the true run lengths as the proportion of censored segments increases.

7.4 Benchmarking against another mouse cerebellum study

Jurrus et al. (2013) address the same problem of finding parallel fibres in the molecular layer of the mouse cerebellum. They also use a similar *classical* stain (where intracellular organelles are stained) and image with a similar scanning electron microscope. We therefore attempt to benchmark our 3D reconstruction performance against that reported in their study. Many of the issues discussed in 6.6.1 are no longer relevant due to the close similarities between the tissue types, staining methods and imaging methods. However, there are still some issues when

comparing results between our work and theirs.

7.4.1 Image z-resolution

The Jurrus et al. study uses an ultramicrotome within the microscope to remove each slice after imaging. As a result, their images have a z-resolution of 50 nm, approximately five times more coarse than the 9.3 nm we achieve by removing each slice with a focussed ion beam. This may make the tracking of parallel fibres in their data more difficult. On the other hand, it may be that the parallel fibres both studies attempt to track may shift sufficiently gradually that this reduced z-resolution does not result in significantly decreased performance. In order to control for this, we could re-evaluate our algorithm using only every 5th slice of our data. Alternatively, if the ground truth used in the Jurrus et al. study becomes available, we could evaluate our algorithm on their test volume.

7.4.2 Level of tracking failure permitted

It is not completely clear from the Jurrus et al. (2013) study whether temporary tracking failures were permitted when determining how many found fibre cross-sections were matched with true fibre cross-sections. Permitting temporary tracking failures results in a significant improvement in the run-length achieved by our algorithm when considering *all matched segments* (section 7.3.5). However, Jurrus et al. appear to only consider the *longest matched segment* for each found fibre and we will do the same here in order to make a fair comparison. For our data, permitting temporary tracking failures when only considering the *longest matched segment* provides only a small improvement in run-length (data not shown). This is because, when considering *all matched segments*, merging segments across regions of temporary tracking failure not only increases the run-length of some segments, but also results in a significant reduction in the *total number* of matched segments. The elimination of these shorter matched segments has a much greater effect on the proportion of segments exceeding each run-length than increasing the run-length of longer segments. When considering only the *longest matched segment* for each found fibre, the number of matched segments considered is always equal to the number of found fibres and does not change when temporary tracking failures are permitted.

7.4.3 Reported run-length measure

The main issue with comparing results between our work and the Jurrus et al. (2013) study is the choice of run-length measure used for comparison. It appears that the Jurrus et al. study only considers the *longest matched segment* for each found fibre when evaluating run-length. It is possible to re-evaluate our results on the same basis by also considering only the longest matched segment for each found fibre. However, the Jurrus et al. algorithm only attempts to

find fibres that exist in *every* slice of their 700×700×70 pixel test volume. As a result, their reported run-length for *fully found fibres* may be equivalent to one of two run-length measures we can generate for our data. The first is the run-length for *fully found fibres* only (those that exist in all 151 slices of our 1,274×852×151 test volume). The second is the run-length for *all found fibres*.

As we can always increase our achieved median run-length for *all found fibres* by enforcing a minimum found fibre length, this second measure will also have multiple possible comparison points. In fact the *fully found fibres* measure is the extreme case of the *all found fibres*, with the minimum found fibre length set to the maximum of 151 slices. Which measure is most appropriate for comparison will therefore depend crucially on the overall proportion of matched cross-sections achieved by the Jurrus et al. study. This in turn will be determined entirely by the number of true fibres present in their test volume, which is not reported. We discuss the two alternative run-length measures below and consider which measure is most appropriate for comparison by estimating the number of true fibres likely to be present in the Jurrus et al. test volume.

Longest matched segment for fully found fibres

One approach to performing an equivalent evaluation to the Jurrus et al. study, is to limit our evaluation of run-length to the subset of our reconstructed fibres that exist in every slice of our test volume and only consider the run-length of the longest *matched segment*. In this case, with no temporary tracking failures permitted, we would find 118 fibres that exist in every slice and 94 (80%) of these would be matched for all 151 slices in our test volume. This is a significantly larger proportion than the 14/56 (25%) reported in the Jurrus et al. study. In figure 7.13 we plot the run-length survival function for our data using this *fully found fibre* measure in green. The performance of our algorithm on this measure clearly out-performs the Jurrus et al. benchmark (dashed black line) by some margin. Note that the green line is identical on all plots, as the parameter sets presented in each plot differ only in the *minimum found fibre length* enforced for the generation of the *all found fibres* run-length measure (blue line). As the minimum found fibre length is set to the maximum of 151 for the *fully found fibre* measure, it is unaffected by this. The dashed black line representing the Jurrus et al. benchmark is also identical on all plots.

Longest matched segment for all fibres

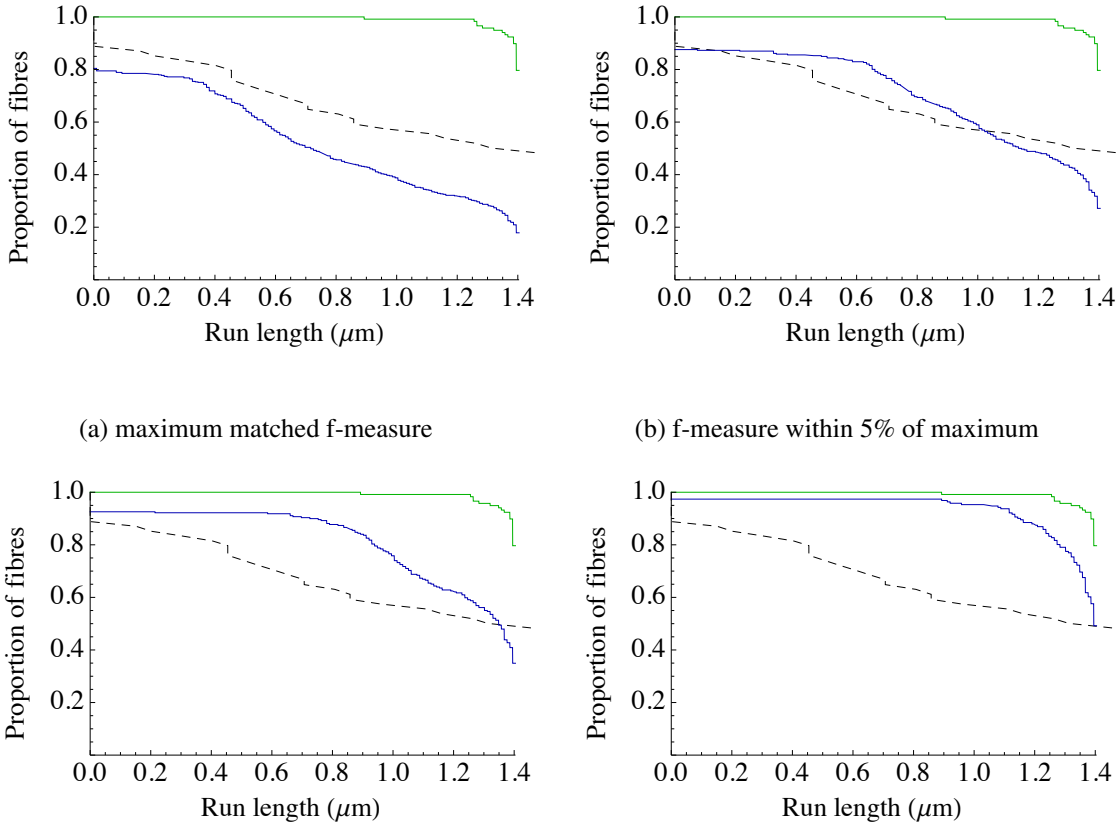
Under the *fully found fibres* measure, only 118/377 (31%) true fibres are matched by *any* found fibre for our data. We would argue that this is an unacceptable price to pay for the excellent run-length achieved by this measure. An alternative is to enforce a lower *minimum found fibre length*, which will ensure more fibres are found at the cost of a lower run-length. In figure 7.13,

we plot run-length survival functions for our *maximum f-measure* parameter set (7.13a) and also parameter sets that maximise run-length while maintaining the f-measure within 5% (7.13b), 10% (7.13c) and 20% (7.13d) of its maximum. As for the *fully found fibre* measure, we consider only the *longest matched segment* for each fibre. The *all found fibres* measure is represented by the blue line on each plot, while the Jurrus et al. benchmark is represented by the dashed black line and is identical on all plots. Our *maximum f-measure* and 5%, 10%, 20% tolerance parameter sets have *minimum found fibre lengths* of 40, 70, 90 and 120 slices respectively.

We estimate the median run-length for the Jurrus et al. data to be $\sim$1.3 μm. The median run lengths for our *maximum f-measure* and 5%, 10%, 20% tolerance parameter sets are 0.72, 1.14, 1.35 and 1.40 μm respectively. Therefore, even when using the *all found fibres* measure, our algorithm out-performs the Jurrus et al. algorithm if a *minimum found fibre length* of 90 slices or greater is enforced. However, it should be noted that our survival function drops off steeply towards the right of the plot for higher f-measure tolerances. In fact, it meets or falls below the Jurrus et al. benchmark by 1.4 μm. The survival function for the Jurrus et al. fibres actually runs for $\sim$3.5 μm in total, with no steep drop-off observed. It might therefore be reasonable to conclude that, considered over the full number of slices in their test volume, the Jurrus et al. algorithm out-performs ours. However, the steep drop-off of our survival function is likely to be at least partially due to *censoring*. Given a finite number of slices in the test volume, some of the *longest matched segments* will touch the edge of the volume at one or both ends. Therefore the length measured for the segment within the test volume is actually a *lower bound* on its true length. The longer a segment, the higher the chance it is censored in the test volume. Therefore, the better an algorithm performs in terms of run-length, the more the run-length survival function will be distorted by censoring effects. The impact of censoring on run-length estimates is discussed in more detail in section 7.3.6. To truly compare the algorithms fairly, we would need to evaluate our algorithm on a test volume that spanned a similar number of slices to the Jurrus et al. study.

Proportion of matched cross-sections

The Jurrus et al. (2013) study does not report the overall number of matched cross-sections. As we can always increase our median *matched segment run-length* at the cost of a reduction in the overall proportion of matched cross-sections (*matched f-measure*), this is crucial for a fair comparison. The fairest comparison would be made between data with equal f-measures. We therefore calculate f-measures for the various run-length results evaluated for our data and attempt to compare these to estimates of the f-measure for the Jurrus et al. data. As we only consider the *longest matched segments* when determining run-lengths, we shall only consider



(a) maximum matched f-measure

(b) f-measure within 5% of maximum

(c) f-measure within 10% of maximum

(d) f-measure within 20% of maximum

Figure 7.13: Benchmarking the performance of our algorithm against that reported in the Jurrus et al. (2013) mouse cerebellum study. We benchmark our results with no permitted tracking failures, using **(a)** our *maximum f-measure* parameters and the parameters that maximise run-length while maintaining an f-measure within **(b)** 5% **(c)** 10% and **(d)** 20% of the maximum. Finally, we benchmark our results for the *longest matched segments* for *fully found fibres* only (**green line**; identical in all plots) and for *all found fibres* (**blue line**). The performance of the Jurrus et al. algorithm is indicated by the **dashed black line** (identical in all plots). It only attempts to find *fully found fibres*. It is not clear what the overall proportion of matched cross-sections in the Jurrus et al. study is. This will determine which of the f-measure tolerance plots represents the fairest comparison. It will also determine whether it is fairest to compare their result for *fully found fibres* against our results for *fully found fibres* (green) or *all found fibres* (blue). Parameter sets for the blue *all found fibre* lines differ only in the *minimum found fibre length* enforced. The green *fully found fibre* line is equivalent to a blue *all found fibre* line with the *minimum found fibre length* set to the maximum of 151.

cross-sections belonging to these segments as *matched* when calculating f-measures.

The *longest matched segment* f-measures for our *maximum f-measure* and 5%, 10%, 20% tolerance parameter sets are 0.79, 0.75, 0.71 and 0.64 respectively, while the f-measure for our *fully found fibres* measure is 0.47. When attempting to estimate the *f-measure* for the Jurrus et al. data we need three pieces of information. Firstly we need to know the total number of *matched cross-sections*. This can be estimated from their reported run length data as ∼1,832. Secondly we need to know the total number of *found fibre cross-sections*. As each of their 56 found fibres exists in all 70 slices of their test volume, this can be caclulated directly to be 3,920. Finally, we need to know the total number of *true fibre cross-sections*. This data is not provided and will critically depend on the number of true fibres present in their test volume. However, by assuming that all true fibres exist in every slice of the test volume, we can directly calculate the total number of *true fibre cross-sections* for different numbers of true fibres. We can then calculate the corresponding f-measures (table 7.1). Comparing these f-measure estimates with those for the various run-length results evaluated for our data, we can establish a reasonable set of comparisons.

True fibres	10	20	**28**	30	40	50	**56**	60	70	80	90	100
F-measure	0.82	0.71	**0.64**	0.63	0.56	0.51	**0.48**	0.47	0.43	0.40	0.37	0.35

Table 7.1: Estimating the *longest matched segment* f-measure for the Jurrus et al. (2013) study. This depends crucially on the number of true fibres present in their test volume. Here we provide f-measure estimates for various numbers of true fibres. Bold entries highlight the estimates when the numbers of true and found fibres are equal (56) and when the number of true fibres is half the number of found fibres (28).

For example, if there are an equal number of true fibres and found fibres in the Jurrus et al. test volume (56), then the corresponding f-measure is likely to be ∼0.48. This is comparable to our *fully found fibre* measure (0.47). Using this measure, our algorithm out-performs the Jurrus et al. algorithm by some margin (green line in figure 7.13). However, the f-measure for the Jurrus et al. data increases as the estimated number of true fibres falls. We should therefore also consider a lower estimate for the number of true fibres present. Assuming the number of true fibres is as low as half the number of found fibres (28), then the Jurrus et al. f-measure rises to 0.64. This is the same as that achieved by our 20% tolerance parameter set, which also comfortably out-performs the Jurrus et al. benchmark across ∼1.4 μm (blue line in figure 7.13d). Depending on the relative impact of *censoring* (section 7.3.6) on the two data sets, our algorithm may or may not continue to out-perform the Jurrus et al. benchmark when evaluated

on a test volume spanning a greater number of slices. Inspecting a sample image from the Bushong and Deerinck (2013) data set that Jurrus et al. used for their study, a 700×700 sample square densely populated with parallel fibres appears to have around 50-60 parallel fibre cross-sections. We would therefore suggest that the *fully found fibre* measure is appropriate to make a fair comparison between the performance of our algorithm and the Jurrus et al. algorithm. On this basis, our algorithm appears to comfortably out-perform theirs. However, this will not be definitively demonstrated until the number of true fibres in the Jurrus et al. test volume is known and we re-evaluate our algorithm using a 50 nm z-resolution on a test volume that is at least 1.4 μm deep. Ideally both algorithms would be evaluated on the same data set. However, the ground truth for the Jurrus et al. test volume is not currently available. We will be publishing both our image and ground truth data once we have completed an analysis of cerebellar ultrastructure using the data. Hopefully the availability of our data set will make it easier to perform precise comparisons across studies in the future.

7.4.4 Summary of benchmark results

We have compared the accuracy of our algorithm against that reported by (Jurrus et al., 2013) on a similar data set. In order to make a fair comparison we have restricted our evaluation to consider only the longest matched segment for each found fibre. However, by varying the *minimum found fibre length* our algorithm can make a trade-off between achieving longer run-lengths for *matched segments* and achieving a higher proportion of matched cross-sections (*f-measure*). Biasing this trade-off to maximise run-length, we outperform the Jurrus et al. benchmark by some margin (green lines in figure 7.13). However, if we bias the trade-off to maximise f-measure instead, we significantly underperform the Jurrus et al. bechmark (blue line in figure 7.13a). In order to establish the level of trade-off that provides the most appropriate comparison, we attempt to estimate the f-measure for the Jurrus et al. algorithm. Examining a sample image from Bushong and Deerinck (2013), we estimate the Jurrus et al. test volume contains 50-60 true fibres, which is approximately the number of fibres found by the Jurrus et al. algorithm (56). Assuming 56 true fibres, we estimate the f-measure for the Jurrus et al. algorithm to be 0.48. This is very close to the 0.47 achieved when we bias the trade-off in our algorithm to maximise run length and consider only *fully found* fibres (green lines in figure 7.13). We would suggest that this is the most appropriate comparison between the two studies. However, even if we halve our estimate of the number of true fibres present, our estimate of the Jurrus et al. f-measure only increases to 0.64 . This is the same as that achieved by our 20% tolerance parameter set, which also comfortably outperforms the Jurrus et al. benchmark (blue line in figure 7.13d). There remain some differences between the data sets used in the two studies and we would ideally

re-evaluate our algorithm on a data set with lower z-resolution and larger z-extent. However, given the margin by which we appear to outperform the Jurrus et al. benchmark on our current data set, we would still expect to perform competitively on this alternative data set. We would therefore claim state of the art 3D reconstruction performance for our algorithm.

7.5 A semi-automated approach

7.5.1 Combining manual and algorithmic inputs

No fully-automated method for reconstructing neurons from electron microscope images currently produces acceptably accurate reconstructions without substantial human proof-reading and correction. A key question is how to minimise the overall amount of human effort required to generate a reconstruction of acceptable quality. Two recent studies have combined human labelling and automated reconstruction in an interesting manner (Briggman, Helmstaedter, and Denk, 2011; Helmstaedter et al., 2013). The automated algorithm generated a volume reconstruction where the reconstructed neurite segments had acceptably accurate boundaries but were relatively short. Humans then provided an independent 1D tracing of the centreline of each neurite of interest. These labelled centrelines joined many short segments of the automated volume reconstruction, threading them like beads on a string. Although each neurite had to be traced by multiple humans to achieve acceptable accuracy, this effort was much lower than that of producing a volume reconstruction via purely manual labelling. The end result was a volume reconstruction that was acceptably accurate over the entire length of each manually traced neurite. This approach is described in Helmstaedter, Briggman, and Denk (2011)

In this work, we explore a different option for enhancing the output of our automated algorithm that also makes use of sparse manual labelling. In our semi-automated scheme, fibre cross-sections are manually labelled in every n^{th} slice by tracing their membrane. It is not necessary to manually assign a unique label to cross-sections from the same fibre across slices. This is left to our algorithm, resulting in a reduction in manual labelling effort. At the labelled slices we base our reconstruction purely on the manual labelling. Between these labelled slices, we blend the information from the manual labelling with algorithm-derived information from the current slice. Specifically, we generate a *ground truth overlap volume* from each manually labelled slice as described in section 5.3.1. We also generate a *predicted overlap volume* from the image at every slice using algorithm 4. At each slice we then blend the local *predicted overlap volume* with the *ground truth overlap volume* from the nearest labelled slice to generate a *composite overlap volume*. This blending is controlled by a *blending co-efficient* (α) that varies with the distance to the nearest labelled slice. To generate the *composite overlap volume*,

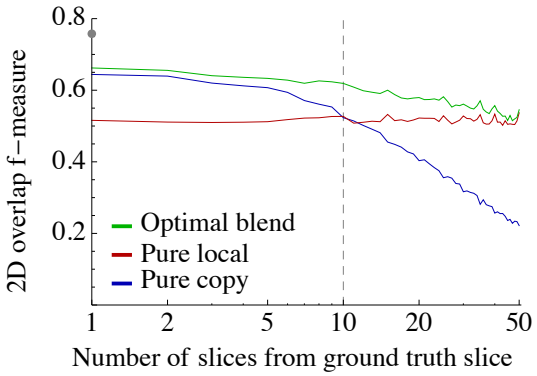
the *ground truth overlap volume* is multiplied by α and the *predicted overlap volume* by (1-α). The *composite overlap volume* is then generated by taking the mean of the weighted ground truth and predicted overlap volumes. This process is described in algorithm 8, and section 7.5.2 describes the derivation of the *optimal blend profile* that determines the optimal α to use as the distance from the nearest labelled slice varies.

Algorithm 8: Combining manual and algorithmic inputs to find 3D tubes

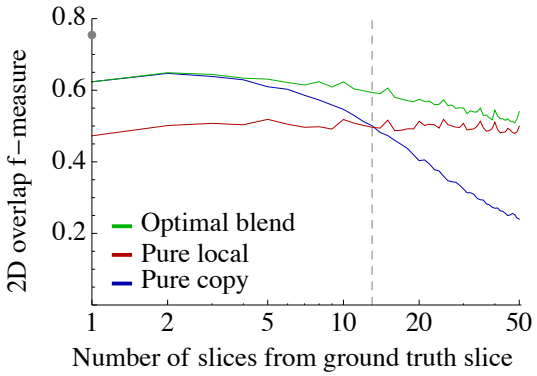
Data: 3D electron microscope image volume; human 2D fibre cross-section labelling at every n^{th} slice; optimal blend profile.
Result: 3D found fibre tubes.
for *all manually labelled slices* **do**
  generate a *ground truth overlap volume* as described in section 5.3.1;
end
for *all slices in volume* **do**
  generate a *predicted overlap volume* from current image slice using algorithm 4;
  get the *ground truth overlap volume* for the nearest labelled slice;
  calculate the *slice separation* between the current slice and the labelled slice;
  set the blend coefficient (α) for this *slice separation* from the *optimal blend profile*;
  weight all values in the *ground truth overlap volume* by multiplying by α;
  weight all values in the *predicted overlap volume* by multiplying by (1-α);
  at each point $\{x,y,r\}$ take the mean of the weighted *ground truth overlap volume* and *predicted overlap volume* to generate the *composite overlap volume*;
  generate predicted circles from the *composite overlap volume* using algorithm 5;
end
generate 3D found fibre tubes from the 2D found fibre circles using algorithm 6;

7.5.2 The spatial influence of manual labelling

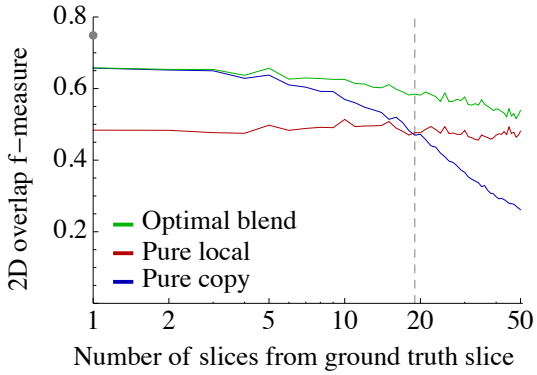
A key question is the proportion with which to blend the *ground truth overlap volume* from the nearest manually labelled slice with the algorithm-derived *predicted overlap volume* from the current slice. This is controlled by the *blending co-efficient* (α). Intuitively, we would expect the usefulness of the ground truth labelling to be high close to the labelled slice and to decrease as the distance from the labelled slice increases. Figure 7.14 illustrates that this is indeed the case. Plots (a)-(c) show the 2D *overlap f-measure* achieved by using only the ground truth from the first slice (blue line), only the local image at the current slice (red line) and the optimal blend between the two (green line). Each plot shows this data for one of three adjacent sub-volumes of our fully labelled volume. Figure 7.14d shows the optimal values for α used to generate the green lines in plots (a)-(c) (dashed grey lines) and the mean *optimal blend profile* (solid black line). We use this *optimal blend profile* in algorithm 8 to determine the optimal value for α given the distance from the nearest labelled slice. The optimal α at a labelled slice is 1.



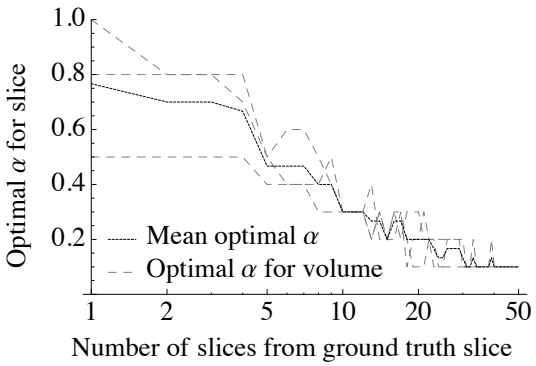
(a) Effect of blending for volume 2161-2211



(b) Effect of blending for volume 2211-2261



(c) Effect of blending for volume 2261-2311



(d) Change in optimal α with distance

Figure 7.14: Spatial influence of manual labelling. **(a)-(c)** The 2D *overlap f-measure* achieved x slices from the start of the volume by using only the local image (**red line**), only the ground truth from the first slice in the volume (**blue line**) and the optimal blend between the two (**green line**). Grey dot on the y-axis indicates the f-measure for the *pure copy* and *optimal blend* at the labelled ground truth slice. Grey dashed lines indicate the slice at which *pure copy* and *pure local* f-measures become equal. The *optimal blend* continues to out-perform both *pure copy* and *pure local* out to at least 50 slices. **(d)** The change in the optimal blending co-efficient (α) as distance from the ground truth slice increases. α is the proportion of ground truth data used. Dashed grey lines show the optimal α for each of the volumes in (a)-(c) and the black solid line shows the per-slice mean across the three volumes. The optimal α at the labelled ground truth slice is 1.

7.5.3 The effect of labelling effort on reconstruction accuracy

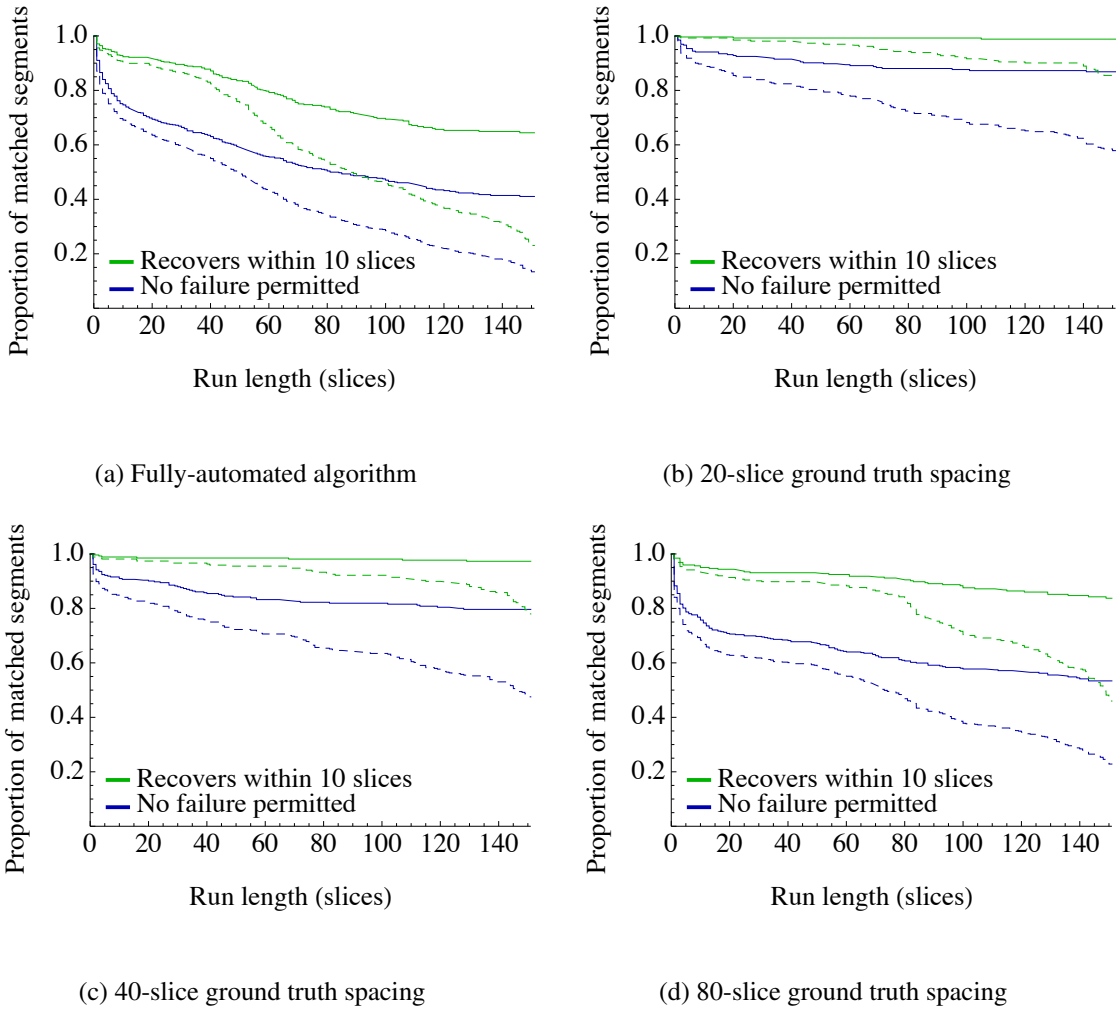
Having established the *optimal blend profile* for combining information from sparse manual labelling and dense algorithm-derived local information, the remaining variable in our scheme is the spacing between manually labelled slices. If this spacing is too sparse we will not achieve an improvement in reconstruction accuracy compared to a fully-automated labelling. If it is too dense, we will not achieve an improvement in reconstruction effort compared to a fully manual labelling. The green lines in figures 7.14 (a)-(c) show the 2D reconstruction performance of our semi-automated approach as the distance to the nearest manually labelled slice increases. The benefit of blending ground truth and algorithm-derived information falls to essentially zero when the nearest manually labelled slice is approximately 50 slices away. This corresponds to a spacing of 100 slices between labelled ground truth slices. In order to determine how this decrease in 2D reconstruction performance impacts 3D reconstruction, we evaluate the effect of providing manually labelled ground truth every 20, 40 and 80 slices (corresponding to distances of 10, 20 and 40 slices to the nearest ground truth slice in figure 7.14).

We evaluate the effect of labelling effort on 3D reconstruction accuracy in three ways. Firstly, we consider the improvement in reconstruction accuracy compared to that achievable via our fully-automated approach. As we measure 3D reconstruction accuracy using two measures, improved accuracy can be considered from two perspectives. Firstly we consider increasing the length of successfully reconstructed segments (*matched segment run-length*), while maintaining the proportion of successfully matched cross-sections (*matched f-measure*). We then consider the reverse, attempting to increase the *matched f-measure* while maintaining the *matched segment run-length*. Finally, we consider the reduction in labelling effort acheivable using a semi-automated approach compared to a purely manual reconstruction while attempting to maintain the same level of reconstruction accuracy.

Increasing run-length while maintaining f-measure at the fully-automated level

One way to utilise the additional information provided by manual labelling is to maximise the median *matched segment run-length*, while maintaining the proportion of matched cross-sections (*matched f-measure*) at the level achieved by our fully-automated approach. We maximise the *unadjusted* median run-length, with no temporary tracking failures permitted and no correction for censoring. Maintaining the f-measure at 0.79, we find that providing manual labelling every 20, 40 and 80 slices increases the run-length from 49 to 151, 147 and 73 respectively. If we examine the corresponding survival functions and consider the effect of permitting temporary tracking failures (section 7.3.5) and correcting for censoring (section 7.3.6), we get a fuller story of the effect of manual labelling effort on run-length. Figure 7.15 illustrates the

improvement in run-length achieved when providing manual ground truth labelling every 20, 40 and 80 slices. Permitting temporary tracking failures of up to 10 slices and accounting for the effect of censoring, 99% of matched segments are reconstructed throughout the full test volume when manually labelled ground truth is provided every 20 slices (7.15b). This proportion falls slightly to 97% when the inter-labelling interval is extended to 40 slices (7.15c) and more substantially to 84% when the interval is further extended to 80 slices (7.15d). All three semi-automated regimes achieve a significant improvement in run-length when compared to the 64% of fibres fully tracked by our fully-automated algorithm (7.15a).



(a) Fully-automated algorithm

(b) 20-slice ground truth spacing

(c) 40-slice ground truth spacing

(d) 80-slice ground truth spacing

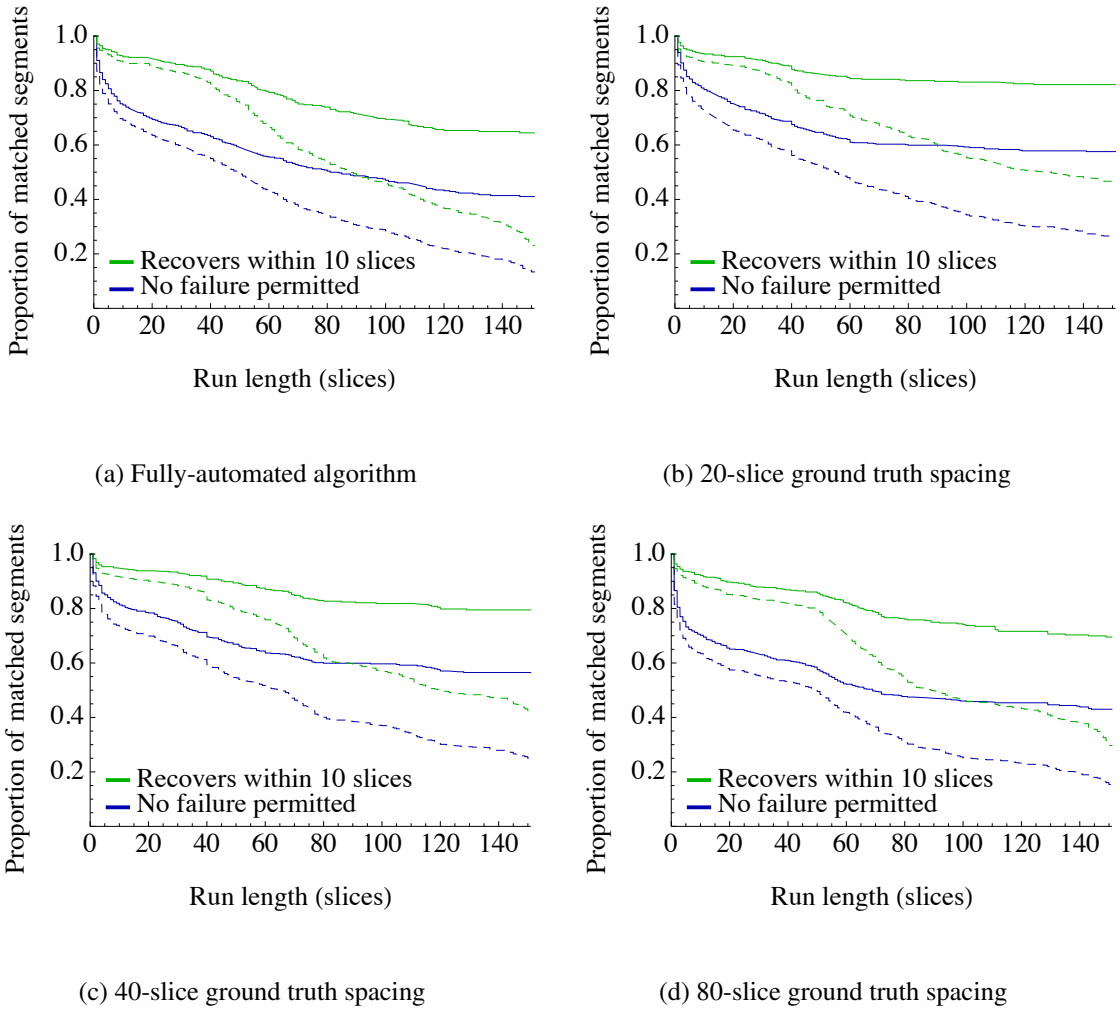
Figure 7.15: Maximising run length with a semi-automated approach. Survival functions for **(a)** our fully-automated algorithm and **(b)-(d)** semi-automated algorithms with manual ground truth labelling provided every 20 slices (b), 40 slices (c) and 80 slices (d). All plots include survival functions with no temporary tracking failures permitted (**blue lines**) and tracking failures of up to 10 slices permitted (**green lines**), both with correction for censoring (**solid lines**) and without (**dashed lines**).

Increasing f-measure while maintaining run-length at the fully-automated level

An alternative approach to utilise the additional information provided by manual labelling is to maximise the proportion of matched cross-sections (*matched f-measure*) while maintaining the median *matched segment run-length* at the level achieved by our fully-automated approach. When maintaining the unadjusted median run-length at $\geq$49 slices, providing manual labelling every 20, 40 and 80 slices increases the achieved f-measure from 0.79 to 0.95, 0.91 and 0.89 respectively. The corresponding unadjusted median run-lengths are 55, 65 and 55. However, if we examine the run-length survival functions after permitting temporary tracking failures of up to 10 slices and correcting for censoring, these correspond to 82%, 79% and 70% of matched segments fully-tracked throughout the test volume. These are all significantly above the 64% achieved by our fully-automated algorithm (figure 7.16). It therefore seems that we can achieve an improvement in the proportion of matched cross-sections from 0.79 to $\sim$ 0.90, while simultaneously increasing the proportion of fully-tracked *matched segments* from 64% to around 80% if we are prepared to provide manual ground truth labelling every 40 slices. This is achieved with a *minimum inter-slice overlap* of 0.2, a *maximum slice separation* of 5 and a *minimum found fibre length* of 40. We use these parameters for our semi-automated approach when making comparisons with a purely manual approach and when visualising reconstructed fibres.

Reducing labelling effort while matching purely manual accuracy

So far we have considered the effect of manual labelling effort on reconstruction accuracy by comparison to the accuracy achieved by our fully-automated algorithm, which requires no manual effort. However, although we manage to achieve significant improvements in reconstruction accuracy compared to our fully-automated algorithm, we still fall short of the "perfect" reconstruction achievable via a purely manual reconstruction. While a fully manual reconstruction can achieve a perfect *matched f-measure* of 1 with all *matched segments* fully tracked, this comes at the cost of significant manual effort. The generation of a purely manually labelled 3D ground truth for our full 2548$\times$852$\times$512 pixel *block03* volume takes approximately 2,500 person hours. In contrast, the generation of a manual 2D labelling every 40 slices would take $\sim$64 person hours. This $\sim$40$\times$ improvement seems like a large reduction in manual effort, and it is comparable to the $\sim$50$\times$ improvement reported by Helmstaedter, Briggman, and Denk (2011) using their centreline tracing scheme. However, the purely manual effort is the cost to achieve a "perfect" reconstruction, while our semi-automated approach will require additional manual effort in order to review and correct its imperfect output. Whether our semi-automated approach is more efficient than a purely manual approach therefore depends critically on the



(a) Fully-automated algorithm

(b) 20-slice ground truth spacing

(c) 40-slice ground truth spacing

(d) 80-slice ground truth spacing

Figure 7.16: Maximising f-measure with a semi-automated approach. Survival functions for **(a)** our fully-automated algorithm and **(b)-(d)** semi-automated algorithms with manual ground truth labelling provided every 20 slices (b), 40 slices (c) and 80 slices (d). All plots include survival functions with no temporary tracking failures permitted (**blue lines**) and tracking failures of up to 10 slices permitted (**green lines**), both with correction for censoring (**solid lines**) and without (**dashed lines**).

effort required to identify and correct the errors in the semi-automated reconstruction. We have not been able to quantify the effort required to perfect the output of our semi-automated approach. However, given that a purely manual reconstruction would take an additional ∼2,400 person hours, we are confident that a proof-read semi-automated reconstruction will still be substantially more efficient than a purely manual one. However, even if no additional manual effort was required for proof-reading, using our semi-automated approach would still only permit the reconstruction of circuits ∼40× the size of those that can be reconstructed with a purely manual approach. While similar improvements in efficiency have facilitated new insights into the local connectivity of cells and microcircuits in the retina (Briggman, Helmstaedter, and Denk, 2011; Helmstaedter et al., 2013), this work still required tens of thousands of hours of labelling effort. Much more significant improvements in reconstruction efficiency will be required if we are to reconstruct the connectivity of larger circuits.

Visualising individual reconstructed fibres

Figure 7.17 visualises the performance of our selected semi-automated algorithm. Ground truth was supplied every 40 slices, and parameters were selected to maximise f-measure while maintaining unadjusted run-length at the level achieved by our fully-automated algorithm. It shows a representative set of true fibres (blue) and their matched found fibres (green and cyan). Fibre 172 is an example of a good long-term one-to-one match (green) that is interrupted by a few cross-sections being better matched by another segment (cyan), which is actually a good long-term match to another true fibre. This is essentially an artefact of the way we match found fibre and true fibre cross-sections independently in each slice. A more sophisticated matching method that considers entire fibres when matching might avoid such issues. Fibre matching is significantly better than that achieved by our full-automated algorithm (figure 7.8), with qualitatively good one-to-one matching of true and found fibres apparent up to fibre 225.

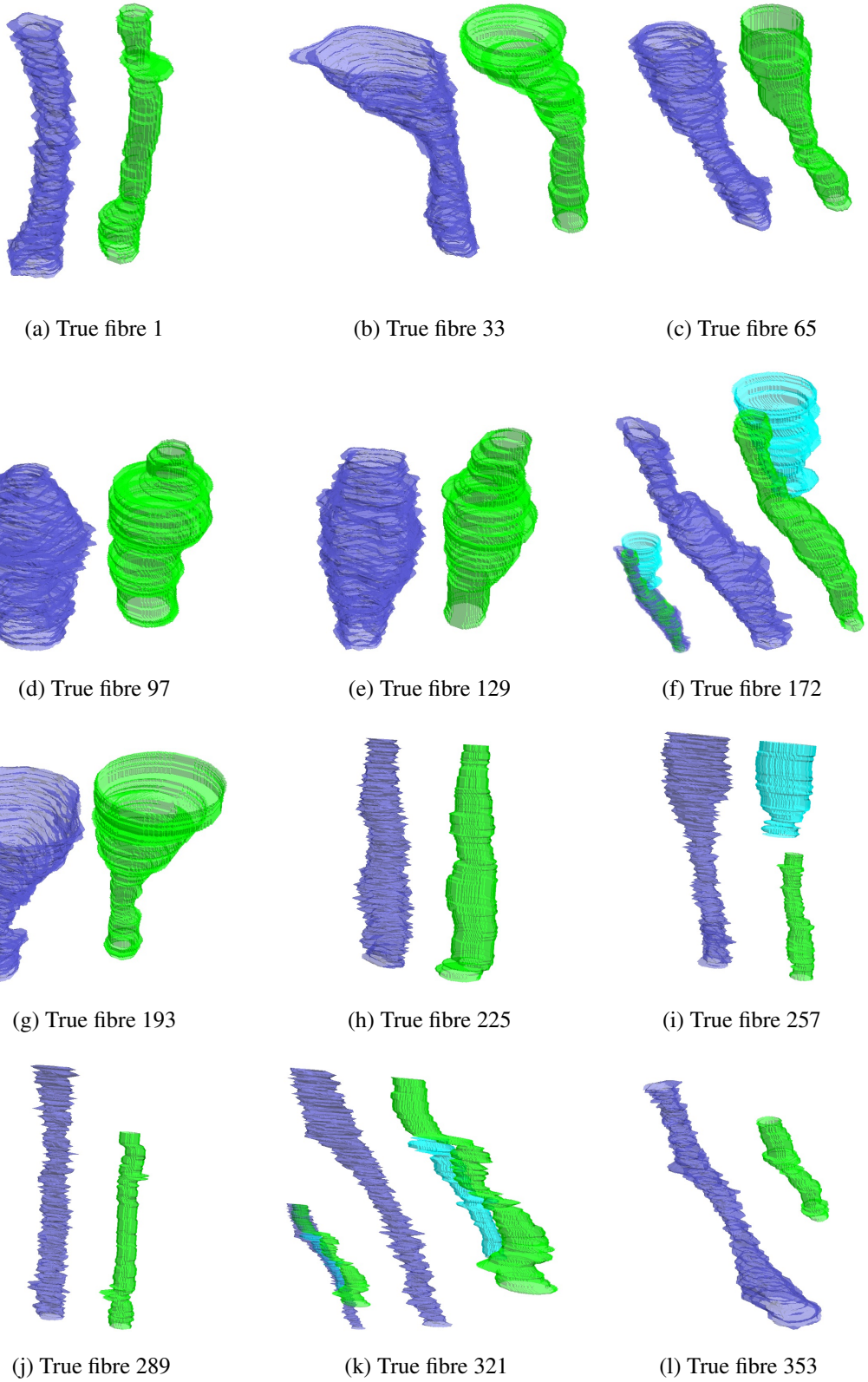
7.6 Limitations and further work

7.6.1 Limitations of our work

There are several limitations of our 3D work, some of which we have already discussed. We discuss their potential impact here, along with the further work required to mitigate them.

Single volume for parameter tuning and evaluation

We use a single sub-volume of our *block03* data set to select the optimum parameters for our 3D algorithm and to evaluate the performance of the algorithm. It is likely that we therefore suffer from over-fitting of our selected parameters to this sub-volume. This means that the 3D reconstruction accuracy we report is likely to be an over-estimate of the accuracy achievable on



(a) True fibre 1 (b) True fibre 33 (c) True fibre 65

(d) True fibre 97 (e) True fibre 129 (f) True fibre 172

(g) True fibre 193 (h) True fibre 225 (i) True fibre 257

(j) True fibre 289 (k) True fibre 321 (l) True fibre 353

Figure 7.17: Example true fibres (**blue**) and their corresponding semi-automated found fibres (**green** and **cyan**). Manual labelling was provided every 40 slices. True and found fibres are separated for clarity. True fibre n corresponds to the n^{th} column in figure 7.6b. Longest found fibres are in green, with additional fibres in cyan. Examples evenly spaced, except true fibre 172. While the green fibre is a good match for true fibre 172, the cyan fibre is a better match for a few cross-sections. It therefore forms a short *matched segment* that splits the green fibre into two *matched segments*. For true fibre 321, an adjacent found fibre incorrectly "snaps" to its top section. Ideally this top section of green fibre would be part of the cyan fibre instead.

a previously unseen data set. To generate an unbiased estimate of 3D reconstruction accuracy, we will need to re-evaluate the performance of our algorithm using a previously unseen sub-volume. However, given that we appear to comfortably out-perform the Jurrus et al. (2013) benchmark, we still expect to do so when using a fully unbiased estimate of reconstruction accuracy.

Use of 3D manual labelling for our semi-automated approach

Our semi-automated approach requires sparse manual labelling of 2D fibre cross-sections. Unfortunately, 2D cross-sections were only classified as fibre or non-fibre for the initial 2D labelling of four slices. When we generated the full 2D manual labelling, we asked the tracers to label all extra-cellular membrane and did not ask them to distinguish between fibres and other cells. We therefore used the fibre/non-fibre classification made after the full 3D labelling to generate a fibre-only subset of the full 2D labelling. It could be argued that this may provide additional information that would not be present were the fibre/non-fibre classification made during a purely 2D reconstruction. We would agree that there likely to be cases where a fibre is unlabelled for a few slices or a non-fibre is labelled for a few slices. However, given that we are bridging gaps of up to 20 slices and discarding found fibres shorter than 40 slices for our fully-automated *maximum f-measure* parameters, it is likely that such local labelling errors would be corrected by our 3D reconstruction algorithm.

Limited depth of our test volume

When correcting for the effects of censoring (section 7.3.6) and when benchmarking our algorithm against the Jurrus et al. (2013) study (section 7.4), we run into issues related to the limited depth of our test volume. As a significant proportion of our matched segments are *right-censored*, our estimates of run-length are lower than those we would achieve with no censoring. While the Kaplan-Meier correction is relatively robust, we would like to re-evaluate our algorithm on a larger sub-set of our *block03* data in order to be fully confident in our run-length estimates. When we benchmark the performance of our algorithm against Jurrus et al., we also experience potential issues with the limited depth of our test volume. Although we appear to comfortably out-perform the benchmark within the 1.4 μm of our test volume, our run-length survival functions fall off steeply toward the end of the volume. As the Jurrus et al. study reports results for almost twice this distance with no steep fall-off, it is possible that their algorithm actually outperforms ours when evaluated over this longer distance. However, the steep drop-off of our survival function is likely to be at least partially due to censoring. The longer a segment, the higher the chance it is censored in the test volume. Therefore, the better an algorithm performs in terms of run-length, the more the run-length survival function will

be distorted by censoring effects. To truly compare the algorithms fairly, we would need to evaluate our algorithm on a test volume that spanned a similar number of slices to the Jurrus et al. study.

7.6.2 Further work

Use of graph based path-finding for 3D reconstruction

The Jurrus et al. (2013) study generates a graph connecting 2D cross-sections in nearby slices, with the edge weights dependent on the cross-correlation between two segmented cross-sections and their relative separation. Found fibres are then reconstructed by applying Dijkstra's algorithm to find the least cost paths through the graph. It would be very interesting to explore the use of a similar approach with our algorithm. The simplest version would replace our fixed *minimum inter-slice overlap* threshold and *maximum slice separation* with edge weight terms that are dependent on the overlap between two found fibre circles and the number of slices separating them. However, as we generate an estimate for the image evidence supporting all possible 2D circles in each slice, we could extend the graph-based approach to replace the current fixed sets of 2D circles with all possible candidate circles in each slice. This could be achieved by adding additional edges and an additional edge weight term that was dependent on the *predicted overlap* our 2D algorithm generates for each candidate circle. In this way we no longer make a hard choice about which circles to place in each 2D slice prior to linking them into 3D tubes. Instead, we incorporate the image evidence for every possible 2D circle in each slice directly into our 3D path finding algorithm. One potential issue with this latter approach might be the computational complexity of path finding on a graph with so many edges. This would need to be evaluated further if such an approach was to be explored.

Chapter 8

Conclusions

8.1 Contributions

In this work we have made four main contributions.

1. A model-based algorithm for reconstructing 2D parallel fibre cross-sections that achieves state of the art 2D reconstruction performance.

2. A fully-automated algorithm for reconstructing 3D parallel fibres that achieves state of the art 3D reconstruction performance.

3. A semi-automated approach for reconstructing 3D parallel fibres that significantly improves reconstruction accuracy compared to our fully-automated approach while requiring $\sim$40$\times$ less labelling effort than a purely manual reconstruction.

4. A "gold standard" ground truth for the molecular layer of the mouse cerebellum that will provide a valuable reference data set for the development and benchmarking of reconstruction algorithms.

We discuss these contributions further below.

8.1.1 A model-based algorithm for 2D reconstruction of fibre cross-sections

We have developed a model-based algorithm for the reconstruction of 2D parallel fibre cross-sections in classically stained electron microscopy images of the cerebellum.

Differences to existing methods

We model the cross-sections of neurites as circles, which addresses the key issues with existing pixel-based and contour-based approaches. We evaluate the image evidence for each circle within an annular region around its perimeter. This results in the consideration of evidence from a larger context than most pixel-based methods, and permits us to integrate image evidence over the entire boundary of a fibre cross-section in a similar manner to contour-based methods.

The use of circles as our model of fibre cross-sections results in a drastic reduction in the number of degrees of freedom compared to contour-based methods. This permits us to evaluate the evidence provided by the image for a full range of candidate circles at each pixel. This exhaustive evaluation of the solution space avoids the problem of local minima associated with contour-based methods.

Key components of the model

In **chapter 5** we introduced the key components of our model. Firstly, we established that circles are a reasonable representation of parallel fibre cross-sections, and that a circle representation is equally useful for generating many models of neural circuit connectivity. Secondly we introduced the concept of the *overlap* of a circle with the manually labelled ground truth as a measure of "fibreness", and established that predicting the *overlap* of each candidate circle with the ground truth is sufficient for reconstructing a suitably accurate reconstruction. Finally we introduced the concept of assessing the image evidence for each circle using the distribution of oriented *Basic Image Features* (oBIFs) within an annular region around its perimeter. We extended the oBIF scheme to include the radial normalisation of oBIF orientations (rBIFs), which is key to achieving sufficiently high reconstruction accuracy.

The 2D reconstruction algorithm

In **chapter 6** we described the incorporation of our circle-based model within a 2D reconstruction algorithm, discussing the selection of algorithm parameters and training data. We explored a range of options for learning to predict the ground truth overlap of a circle from its associated rBIF histogram, and determined that logistic regression performs as well as more sophisticated techniques when square-rooted histograms are used. Finally, we evaluated the effectiveness of our 2D algorithm by benchmarking it against *ilastik*, a state of the art pixel-based classifier. The performance of our algorithm and *ilastik* are very similar, achieving $\sim$50% on an overlap-based f-measure. We would therefore claim state of the art performance at reconstructing 2D parallel fibre cross-sections.

8.1.2 A fully-automated algorithm for reconstructing 3D parallel fibres

In **chapter 7** we extended our algorithm for finding fibre cross-sections in 2D to reconstruct fibres in 3D by linking cross-sections across slices. We introduced a pair of measures to capture the quality of a 3D reconstruction, setting them in the context of other measures used in the literature. We characterised the effects of parameter selection on algorithm performance, and discussed the trade-off between maximising the length of successfully reconstructed fibre segments and maximising the overall proportion of fibre cross-sections that are well found. Fi-

nally, we benchmarked our algorithm against a recently published study addressing the same reconstruction problem on a similar data set. After re-evaluating the accuracy of our algorithm on the same basis, and making some conservative estimates of the overall proportion of well found fibre cross-sections in this study, we appear to outperform the benchmark by a reasonable margin. We would therefore claim state of the art performance at reconstructing parallel fibres in 3D.

8.1.3 A semi-automated approach for reconstructing 3D parallel fibres

In **chapter 7** we also extended our 3D algorithm further to incorporate information from a sparse 2D manual membrane labelling. This semi-automated approach resulted in a significant improvement in reconstruction accuracy compared to our fully-automated algorithm. Around 90% of fibre cross-sections are well found, and ∼80% of fibres are tracked successfully throughout the entire test volume. This semi-automated approach requires ∼40× less manual labelling effort than a purely manual labelling approach. This reduction in labelling effort is comparable to the ∼50× reduction recently reported for a semi-automated approach using manually labelled neurite centrelines. We have yet to quantify the additional manual effort required to correct the remaining errors in our semi-automated reconstruction, and this will reduce our final achieved efficiency gain. However, we would expect our fully-corrected semi-automated approach to remain significantly more efficient than a purely manual approach.

8.1.4 A "gold standard" ground truth for the mouse cerebellum

We have generated a high quality 3D ground truth labelling for a region of mouse cerebellum. Once we have published an analysis of cerebellar ultrastructure using this data, we will publish both the electron microscope images and the ground truth labelling in the open access Cell Centered Database (CCDB). Our data set is significantly larger than those already published, and will provide a valuable reference data set for the development and benchmarking of neural reconstruction algorithms. In particular, our data set is uniquely suitable for benchmarking 2+1D approaches to neural reconstruction.

8.2 Issues and limitations

8.2.1 Restriction to 2+1D problems

Our algorithm takes a 2+1D approach to reconstructing fibres in 3D. We first find fibre cross-sections in 2D slices, and then link these cross-sections across slices to form 3D fibres. This approach works well in the molecular layer of the cerebellum, which can be imaged such that the majority of fibres run almost perpendicular to the image plane, but will not transfer well to other areas of the brain where fibres can have a wide range of orientations. However, the

parallel fibres we reconstruct are the sole output of the cerebellar granule cells, which comprise ∼80% of the neurons in the brain. Therefore, even if 2+1D approaches such as ours are limited to reconstructing only parallel fibres, they can still be extremely useful.

8.2.2 Benchmarking difficulties

When evaluating both our 2D and 3D algorithms, the lack of a suitable publicly available benchmark data set for 2+1D approaches made comparison with the results of other studies difficult. For our 2D algorithm, we struggled to find results reported against a similar data set, as the Jurrus et al. (2013) study was not published at the time this 2D work was done. We therefore took a published version of a state of the art algorithm and evaluated it on our data set. However, differences in the natural density of the 2D reconstructions made by the two algorithms meant it was still difficult to ensure a valid comparison. When we came to benchmark our 3D algorithm, the Jurrus et al. study had been published. This addressed the same reconstruction problem in a similar data set, making a cross-study benchmark feasible. However, it was still difficult to ensure a valid comparison. This was due to differences in the z-resolution and z-extent of the two data sets, and because some of the information required to generate a comparable performance measure was not available. We plan to address these issues in future work. Although we believe we have made valid comparisons with both our 2D and 3D benchmarking, this process would have been much easier if there was an existing benchmark for 2+1D approaches. We will be providing such a benchmarking resource to the community by publishing our data set.

8.3 Future work

8.3.1 Refined 3D benchmarking

In order to provide a closer match for the Jurrus et al. (2013) data set, we plan to down-sample our image data in the z-dimension and increase the z-extent we use for evaluation. We will then re-evaluate our algorithm to confirm that it still outperforms the benchmark.

8.3.2 A neuroscientific analysis of our ground truth reconstruction

Although we primarily generated our 3D ground truth labelling to support the development and evaluation of our reconstruction algorithms, there is an opportunity to use it to analyse the ultrastructure of the cerebellar molecular layer. We plan to perform this analysis before publishing our data set.

8.3.3 Tube finding in three dimensions

Although we were unable to extend rBIFs to 3D during the course of this work, this remains a promising avenue for exploration. In 3D, we would replace circles with short 3D tube segments

and normalise rBIF orientations to be relative to the vector from a pixel to the tube centreline. We would then find the 3D tube segments that best locally represent a fibre and link them together to form full 3D fibres. However, if we continue to perform an exhaustive search for the best supported fibre representations, moving from circles to tubes will significantly increase the computational burden of our algorithm. Some further analysis will be required to determine whether a 3D tube approach can be achieved with reasonable computational resources.

8.3.4 Combining different reconstruction methods

While we were unable to find a useful method of combining our 2D algorithm with the *ilastik* algorithm we were benchmarking against, we believe there should be a way to do so. Further research may be able to discover a useful method for combining the output of these two algorithms. The graph-based approach used by Jurrus et al. (2013) to link cross-sections across slices could also be usefully applied to our approach. We currently independently select the best supported circles in each slice before linking them across slices. The graph-based approach may permit us to link these independently chosen circles together better. However, as we maintain a "fibreness" score for every possible circle in each slice, the graph-based approach may also permit the selection of the set of circles that form the best globally supported 3D fibres across all slices simultaneously.

Appendix A

Sample preparation and imaging

Acknowledgement: The following sections on sample preparation (A.1) and image acquisition (A.2) were kindly provided by Sarah Rieubland and are not the work of the thesis author. They are provided here to ensure that all information required to reproduce the type of EM image data this thesis relies on is present in the thesis.

A.1 Sample preparation

Six week old adult mice, anesthetized with ketamine/xylazine, underwent fixation by cardiac perfusion of 50 mL of phosphate buffer followed by 50 mL 2.5% glutaraldehyde and 2% paraformaldehyde in phosphate buffer (0.1 M, pH 7.4). After perfusion, the brain was removed and immersed in fixative solution for at least 20 minutes. The cerebellum was dissected and the cerebellar vermis isolated by two parasagittal cuts. 100 μm thick sagittal sections were cut with a vibratome slicer and then processed for EM preparation. The slices were postfixed for 30 minutes in 1% osmium tetroxide and 1.5% potassium ferrocyanide in 0.1M sodium cacodylate buffer, then immersed in 1% thiocarbohydrazide (TCH) solution, and finally stained a second time with 1% osmium tetroxide in 0.1M sodium cacodylate buffer. Slices were then dehydrated through an ascending concentration of ethanol. Propylene oxide was used to progressively infiltrate the slices with resin. Durcupan resin (Fluka) was prepared from the four components A: 10 mL, B: 10 mL, C: 0.35 mL, D: 0.15 mL. Infiltrated slices were embedded flat between glass slides and coverslips and put in the oven for 48 hours at 60°C. The resin sections were glued at the tip of a resin block and the top and corner surfaces were polished using a freshly cut glass knife in an ultramicrotome (Leica EM UC6). Samples were then mounted on metal stubs, covered with silver paint (Agar) and sputtered with gold before being loaded in the electron microscope.

A.2 Image acquisition

Samples were loaded in the focused ion beam electron microscope (FIBSEM, NVision 40, Zeiss) and their orientation adjusted ($\sim$54°) to polish the front surface with the ion beam and image with the electron beam at a 36° angle. Low energy electrons (accelerating potential = 1.5 keV) were used to minimise the interaction volume. Back-scattered electrons were detected with the ESB detector and the voltage acceptance, brightness and contrast were adjusted to optimise the image contrast and signal to noise (voltage acceptance = 300V; brightness = 0%; contrast = 57%). Slow scan speed (setting 9; dwell time = 25 μs/pixel) and a 120 μm aperture provided high resolution images, and were combined with FIB milling with low probe current (I_{probe} = 1.5 — 3 nA) to obtain smooth polishing and regular sectioning of the sample. Targeted regions of the cerebellar molecular layer were imaged in sagittal sections with isotropic resolution of 9.3 nm.

A.3 Image post-processing

The acquired images were registered into a common reference frame using the *Linear Stack Alignment with SIFT* plug-in for Fiji (SIFT: Lowe, 2004; Fiji: Schindelin et al., 2012). A 2548x852x512 voxel sub-volume was used for this work. The Häusser lab identifier for the full image volume is *Roth::100213_16_R-OTO*. The lab identifier for the cropped sub-volume is *OReilly::block03*. SIFT alignment parameters and sub-volume offsets are provided below.

SIFT alignment parameters

- Scale Invariant Interest Point Detector: *initial gaussian blur = 1.60 px; steps per scale octave = 3; minimum image size = 64; maximum image size = 1024.*
- Feature Descriptor: *feature descriptor size = 4; feature descriptor orientation bins = 8; closest/next closest ratio = 0.92.*
- Geometric Consensus Filter: *maximal alignment error = 25.00 px; inlier ratio = 0.05; expected transformation = Rigid;*
- Output: *interpolate = checked; show info = unchecked.*

Sub-volume offsets

- x-offset = +428 (i.e. co-ordinates in full volume are co-ordinates in sub-volume + 428)
- y-offset = +1400 (i.e. co-ordinates in full volume are co-ordinates in sub-volume + 1400)
- z-offset = +941 (i.e. co-ordinates in full volume are co-ordinates in sub-volume + 941)

Appendix B

Jaccard index publication history

Jaccard first proposed his now famous similarity index in a 1901 edition of the french language bulletin of the provincial *Société vaudoise des sciences naturelles*, in an paper entitled *Distribution de la flore alpine dans le Bassin des Drouces et dans quelques regions voisines* (pp. 241-272: Jaccard, 1901a). In this paper Jaccard refers to his measure as the *coefficient de communauté* (or *coefficient of community*) and defines it as the number of species common to two regions divided by the total number of species across the two regions (see footnote 1 on p.249). We present a transcription of the original definition from this paper below, along with a translation into english (courtesy of Google Translate and some rusty high school french).

Original french

> [1] *Pour évaluer la proportion d'espèces communes, il suffit de soustraire du total des deux listes compardes, le nombre des esteèces communes. Ainsi Triente 470 + W. 350 = 820. 820 - 295 esp. communes = 525 esp. differéntes dont 295 sont communes aux deus listes soit plus de la moitié,* $\frac{56}{100}$ *environs.*

English translation

> [1] *To evaluate the proportion of common species, it is sufficient to subtract from the total of the two compared lists the number of common species. Therefore Triente 470 + W. 350 = 820. 820 - 295 common species = 525 unique species of which 295, or more than half, are common to both lists (approximately* $\frac{56}{100}$*).*

Interestingly, Jaccard authored a second paper in the same 1901 bulletin edition. This paper was entitled *Étude comparative de la distribution florale dans une portion des Alpes et des Jura* (pp. 547-579: Jaccard, 1901b) and Jaccard makes use of his newly defined *coefficient de communauté* in his analysis. This has resulted in several articles incorrectly citing this second 1901 paper as the source of the Jaccard index even though the measure is not defined in it. The *coefficient de communauté* values referred to in the text are contained in an unlabelled column

across two tables and are calculated by combining data in these tables with data from a third. Without existing knowledge of the measure's definition it is non-trivial to reverse engineer it from the information provided in this second paper alone. Out of all the papers discussed, this is easily the poorest candidate for use as a reference for the Jaccard index.

To add to the confusion caused by the two 1901 papers, a revised version of Jaccard's original *Distribution de la flore alpine dans le Bassin des Drouces et dans quelques regions voisines* paper appeared in a 1907 edition of the Paris journal *Revue générale des Sciences pures et appliquées* under the title *La distribution de la flore dans la zone Alpine* (Jaccard, 1907). This paper included the definition of the *coefficient de communuaté* and was later translated in a 1912 edition of *New Phytologist* (Jaccard, 1912), providing the earliest english language reference for the Jaccard index. Finally, Jaccard again included the definition for his *coefficient de communuaté* in a 1908 paper for the bulletin of the *Société vaudoise des sciences naturelles* entitled *Nouvelles recherches sur la distribution florale* (Jaccard, 1908), providing yet another alternative reference for the Jaccard index.

We would suggest that the most appropriate reference for the Jaccard index is the 1901 *Distribution de la flore alpine dans le Bassin des Drouces et dans quelques regions voisines* paper (Jaccard, 1901a), as this is where Jaccard originally introduces the measure. If an english language reference is required then the 1912 *New Phytology* paper is a reasonable choice, as it translates a revised version of the original 1901 paper.

Appendix C

Generating BIFs

In this appendix we provide algorithms for generating BIFs and quantising their orientation.

Algorithm 9: Generating a BIF map for an image

Data: 2D electron microscope image
Result: BIF class (and orientation for oBIFs and rBIFs) at each pixel in image
generate a second-order family of Derivative of Gaussian filters (table 5.2), using a standard deviation (σ) of 1.75;
convolve the image separately with each filter to generate filtered images L_{00}, L_{10}, L_{01}, L_{20}, L_{11} and L_{02}. Mirror border pixels to ensure filtered images have the same size as the original image;
for *all pixels in image* **do**
    calculate the response magnitude (R) for each of the seven BIF classes by combining the filtered L_{nn} images as detailed in table 5.1;
    assign the pixel to the BIF class with the largest response magnitude (R);
    if *bif type is oBIF or rBIF* **then**
        if *bif class is gradient* **then**
            set oBIF orientation unit vector (v) to $v_x = \frac{L_{10}}{\sqrt{L_{10}^2+L_{01}^2}}$; $v_y = \frac{L_{01}}{\sqrt{L_{10}^2+L_{01}^2}}$
        end
        if *bif class is line or saddle* **then**
            set oBIF orientation unit vector (v) to the eigenvector associated with the smallest eigenvalue of the *Hessian matrix*, $H = \begin{bmatrix} L_{20} & L_{11} \\ L_{11} & L_{02} \end{bmatrix}$;
        end
    end
end

Algorithm 10: Quantising oriented BIFs (oBIFs and rBIFs)

Data: Non-quantised oBIFs or rBIFs
Result: Quantised oBIFs or rBIFs
if *bif class is gradient* **then**
  for θ_Q *in* $0, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}, \pi, \frac{5\pi}{4}, \frac{6\pi}{4}, \frac{7\pi}{4}$ **do**
    set the oBIF orientation angle θ to atan2(v_y,v_x);
    assign a fractional pixel to the θ_Q gradient histogram bin according to $fraction = \frac{1}{\sqrt{2\pi\sigma}} \exp \frac{(\theta-\theta_Q)^2}{2\sigma^2}$, where σ is set to 0.65;
  end
end
if *bif class is line or saddle* **then**
  for θ_Q *in* $0, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}$ **do**
    set the oBIF orientation angle θ to atan2(v_y,v_x);
    assign a fractional pixel to the θ_Q gradient histogram bin according to $fraction = max\left(\frac{1}{\sqrt{2\pi\sigma}} \exp \frac{(\theta-\theta_Q)^2}{2\sigma^2}, \frac{1}{\sqrt{2\pi\sigma}} \exp \frac{(\theta-\theta_Q+\pi)^2}{2\sigma^2}\right)$, where σ is set to 0.65;
  end
end

Bibliography

Adelson, Edward H. and James R. Bergen (1985). "Spatiotemporal energy models for the perception of motion". In: *Journal of the Optical Society of America A, Optics and Image Science* 2.2, pp. 284–299. DOI: 10.1364/JOSAA.2.000284.

Anderberg, Michael R. (1973). *Cluster analysis for applications*. Academin Press, New York.

Anderson, James R et al. (2009). "A Computational Framework for Ultrastructural Mapping of Neural Circuitry". In: *PLoS Biology* 7.3, e1000074–. DOI: 10.1371/journal.pbio.1000074.

Andres, Björn et al. (2008). "Segmentation of SBFSEM Volume Data of Neural Tissue by Hierarchical Classification". In: *Pattern Recognition (Proceedings of the 30th DAGM Symposium)*. Vol. LNCS 5096. Lecture Notes in Computer Science, pp. 142–152. DOI: 10.1007/978-3-540-69321-5_15.

Andres, Björn et al. (2012). "3D segmentation of SBFSEM images of neuropil by a graphical model over supervoxel boundaries". In: *Medical Image Analysis* 16.4, pp. 796 –805. ISSN: 1361-8415. DOI: 10.1016/j.media.2011.11.004.

Azevedo, Frederico A.C. et al. (2009). "Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain". In: *Jornal of Comparative Neurololgy* 513.5, pp. 532–541. ISSN: 1096-9861. DOI: 10.1002/cne.21974.

Beucher, Serge and Christian Lantuéjoul (1979). "Use of watersheds in contour detection". In: *International workshop on image processing, real-time edge and motion detection*. eprint: http://cmm.ensmp.fr/~beucher/publi/watershed.pdf.

Bock, Davi D. et al. (2011). "Network anatomy and in vivo physiology of visual cortical neurons". In: *Nature* 471.7337, pp. 177–182. ISSN: 0028-0836. DOI: 10.1038/nature09802.

Borton, David A et al. (2013). "An implantable wireless neural interface for recording cortical circuit dynamics in moving primates". In: *Journal of Neural Engineering* 10.2, p. 026010. DOI: 10.1088/1741-2560/10/2/026010.

Breiman, Leo (1996). "Bagging predictors". English. In: *Machine Learning* 24.2, pp. 123–140. ISSN: 0885-6125. DOI: 10.1007/BF00058655.

Breiman, Leo (2001). "Random Forests". English. In: *Machine Learning* 45.1, pp. 5–32. ISSN: 0885-6125. DOI: 10.1023/A:1010933404324.

Breiman, Leo et al. (1984). *Classification and regression trees*. Wadsworth International Group: Belmont, CA.

Briggman, Kevin L., Moritz Helmstaedter, and Winfried Denk (2011). "Wiring specificity in the direction-selectivity circuit of the retina". In: *Nature* 471.7337, pp. 183–188. ISSN: 0028-0836. DOI: 10.1038/nature09818.

Bushong, Eric and Tom Deerinck (2013). *Cell Centered Database: Microscopy product ID: 8192*. eprint: http://ccdb.ucsd.edu/sand/main?mpid=8192&event=displaySum.

Cajal, Santiago Ramon Y (1894). "The Croonian Lecture: La Fine Structure des Centres Nerveux". In: *Proceedings of the Royal Society of London* 55.331-335, pp. 444–468. DOI: 10.1098/rspl.1894.0063.

Cardona, Albert (2012). *ISBI 2012 Challenge Website: Segmentation of neuronal structures in EM stacks*. eprint: http://fiji.sc/Segmentation_of_neuronal_structures_in_EM_stacks_challenge_-_ISBI_2012.

Cardona, Albert et al. (2009). "Drosophila Brain Development: Closing the Gap between a Macroarchitectural and Microarchitectural Approach". In: *Cold Spring Harbor Symposia on Quantitative Biology* 74, pp. 235–248. DOI: 10.1101/sqb.2009.74.037.

Carnevale, N.T. and M.L. Hines (2006). *The NEURON Book*. Cambridge University Press.

Chen, Beth L., David H. Hall, and Dmitri B. Chklovskii (2006). "Wiring optimization can relate neuronal structure and function". In: *Proceedings of the National Academy of Sciences of the United States of America* 103.12, pp. 4723–4728. DOI: 10.1073/pnas.0506806103.

Chklovskii, Dmitri B., Shiv Vitaladevuni, and Louis K Scheffer (2010). "Semi-automated reconstruction of neural circuits using electron microscopy". In: *Current Opinion in Neurobiology* 20.5, pp. 667 –675. ISSN: 0959-4388. DOI: 10.1016/j.conb.2010.08.002.

Cireşan, Dan Claudiu et al. (2010). "Deep, Big, Simple Neural Nets for Handwritten Digit Recognition". In: *Neural Computation* 22.12, pp. 3207–3220. ISSN: 0899-7667. DOI: 10.1162/NECO_a_00052.

Cireşan, Dan et al. (2012). "Deep Neural Networks Segment Neuronal Membranes in Electron Microscopy Images". In: *Conference of the Neural Information Processing Systems Foundation (NIPS)*. eprint: http://www.idsia.ch/~juergen/nips2012.pdf.

Consortium, International Human Genome Sequencing et al. (2001). "Initial sequencing and analysis of the human genome". In: *Nature* 409.6822, pp. 860–921. DOI: 10.1038/35057062.

Consortium, The 1000 Genomes Project (2012). "An integrated map of genetic variation from 1,092 human genomes". In: *Nature* 491, pp. 56–65. DOI: 10.1038/nature11632.

Consortium, The C.elegans Sequencing (1998). "Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology". In: *Science* 282.5396, pp. 2012–2018. DOI: 10.1126/science.282.5396.2012.

Cooper, Seth et al. (2010). "Predicting protein structures with a multiplayer online game". In: *Nature* 466.7307, pp. 756–760. ISSN: 0028-0836. DOI: 10.1038/nature09304.

Crosier, Mike and Lewis D. Griffin (2010). "Using Basic Image Features for Texture Classification". In: *International Journal of Computer Vision* 88 (3), pp. 447–460. ISSN: 0920-5691. DOI: 10.1007/s11263-009-0315-0.

Denk, Winfried and Heinz Horstmann (2004). "Serial Block-Face Scanning Electron Microscopy to Reconstruct Three-Dimensional Tissue Nanostructure". In: *PLoS Biology* 2.11, e329. DOI: 10.1371/journal.pbio.0020329.

Dice, Lee R. (1945). "Measures of the Amount of Ecologic Association Between Species". In: *Ecology* 26.3, pp. 297–302. ISSN: 00129658. DOI: 10.2307/1932409.

DoE (1991). *Understanding Our Genetic Inheritance: The U.S. Human Genome Project, The First Five Years, FY 1991-1995*. Tech. rep. U.S. Department of Energy. eprint: http://web.ornl.gov/sci/techresources/Human_Genome/project/5yrplan/index.shtml.

Durbin, Richard (1987). "Studies on the development and organisation of the nervous system of Caenorhabditis Elegans". PhD thesis. Kings College, Cambridge. eprint: http://www.wormatlas.org/ver1/durbinv1.2/durbinindex.html.

Einevoll, Gaute T et al. (2012). "Towards reliable spike-train recordings from thousands of neurons with multielectrodes". In: *Current Opinion in Neurobiology* 22.1, pp. 11 –17. ISSN: 0959-4388. DOI: 10.1016/j.conb.2011.10.001.

Emerson, Robert C., James R. Bergen, and Edward H. Adelson (1992). "Directionally selective complex cells and the computation of motion energy in cat visual cortex". In: *Vision Research* 32.2, pp. 203–218. ISSN: 0042-6989. DOI: 10.1016/0042-6989(92)90130-B.

Essen, David C. Van et al. (2013). "The WU-Minn Human Connectome Project: An overview". In: *NeuroImage* 80, pp. 62 –79. ISSN: 1053-8119. DOI: http://dx.doi.org/10.1016/j.neuroimage.2013.05.041.

Ethier, C. et al. (2012). "Restoration of grasp following paralysis through brain-controlled stimulation of muscles". In: *Nature* advance online publication, ISSN: 1476-4687. DOI: 10.1038/nature10987.

Felleman, Daniel J. and David C. Van Essen (1991). "Distributed Hierarchical Processing in the Primate Cerebral Cortex". In: *Cerebral Cortex* 1.1, pp. 1–47. DOI: 10.1093/cercor/1.1.1.

Field, Greg D. et al. (2010). "Functional connectivity in the retina at the resolution of photoreceptors". In: *Nature* 467.7316, pp. 673–677. ISSN: 0028-0836. DOI: 10.1038/nature09424.

Freeman, W.T. and Edward H. Adelson (1991). "The design and use of steerable filters". In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 13.9, pp. 891–906. ISSN: 0162-8828. DOI: 10.1109/34.93808.

Fukushima, Kunihiko (1980). "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position". English. In: *Biological Cybernetics* 36.4, pp. 193–202. ISSN: 0340-1200. DOI: 10.1007/BF00344251.

Galvani, Luigi Aloysii (1791). "De viribus electricitatis in motu musculari". In: *De Bononiensi scientiarum et artium instituto atque academia commentarii* 7, pp. 363–418. eprint: http://steling.alternativaverde.it/documenti/Galvani/DeViribus.pdf.

— (1792). *De viribus electricitatis in motu musculari*. Ed. by Giovanni Aldini. Apud societatem typographicam. eprint: http://archive.org/details/aloysiigalvaniin00galv.

Giuly, Richard, Maryann Martone, and Mark Ellisman (2012). "Method: automatic segmentation of mitochondria utilizing patch classification, contour pair classification, and automatically seeded level sets". In: *BMC Bioinformatics* 13.1, p. 29. ISSN: 1471-2105. DOI: 10.1186/1471-2105-13-29.

Glasner, Daniel et al. (2011). "High Resolution Segmentation of Neuronal Tissues from Low Depth-Resolution EM Imagery". In: *Energy Minimization Methods in Computer Vision and Pattern Recognition*. Ed. by Yuri Boykov et al. Vol. 6819. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 261–272. ISBN: 978-3-642-23093-6. DOI: 10.1007/978-3-642-23094-3_19.

Gleeson, Padraig, Volker Steuber, and R. Angus Silver (2007). "neuroConstruct: A Tool for Modeling Networks of Neurons in 3D Space". In: *Neurotechnique* 54.2, pp. 219–235. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2007.03.025.

Goldschmidt, RB. (1908). "Das nervensystem von Ascaris lumbricoides und megalocephala, I". In: *Zeitschrift für wissenschaftliche Zoologie* 90, pp. 73–136. eprint: http://www.biodiversitylibrary.org/item/38745#page/87/mode/1up.

— (1909). "Das nervensystem von Ascaris lumbricoides und megalocephala, I". In: *Zeitschrift für wissenschaftliche Zoologie* 92, pp. 306–357. eprint: http://www.biodiversitylibrary.org/item/38775#page/316/mode/1up.

Green, Robert Montraville (1953). *Galvani on Electricity*. Elizabeth Light, Cambridge MA. eprint: http://archive.org/details/commentaryonthee002243mbp.

Gustafsson, Mats G.L. et al. (2008). "Three-Dimensional Resolution Doubling in Wide-Field Fluorescence Microscopy by Structured Illumination". In: *Biophysical Journal* 94.12, pp. 4957 –4970. ISSN: 0006-3495. DOI: 10.1529/biophysj.107.120345.

Hagmann, Patric (2005). "From diffusion MRI to brain connectomics". PhD thesis. École polytechnique fédérale de Lausanne. DOI: 10.5075/epfl-thesis-3230.

Hayworth, Kenneth J et al. (2007). "Automatic collection of ultrathin brain sections for large volume neuronal circuit reconstruction (poster)".

Hayworth, K.J. et al. (2010). "Divide and conquer - Lossless thick sectioning of plastic-embedded brain tissue to parallelize large volume serial reconstructions". In: *Society for Neuroscience Annual Meeting (SfN). Program No. 516.9. Poster No. PPP4.*

Helmstaedter, Moritz, Kevin L. Briggman, and Winfried Denk (2011). "High-accuracy neurite reconstruction for high-throughput neuroanatomy". In: *Nature Neuroscience* 14.8, pp. 1081–1088. ISSN: 1097-6256. DOI: 10.1038/nn.2868.

Helmstaedter, Moritz et al. (2013). "Connectomic reconstruction of the inner plexiform layer in the mouse retina". In: *Nature* 500.7461, pp. 168–174. ISSN: 0028-0836. eprint: http://dx.doi.org/10.1038/nature12346.

Hodgkin, A. L. and A. F. Huxley (1939). "Action Potentials Recorded from Inside a Nerve Fibre". In: *Nature* 144, pp. 710–711. DOI: 10.1038/144710a0.

Ho, Tin Kam (1998). "The random subspace method for constructing decision forests". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* 20.8, pp. 832–844. ISSN: 0162-8828. DOI: 10.1109/34.709601.

Hubel, D. H. and T. N. Wiesel (1959). "Receptive fields of single neurones in the cat's striate cortex". In: *The Journal of Physiology* 148.3, pp. 574–591. eprint: http://jp.physoc.org/content/148/3/574.short.

— (1962). "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex". In: *The Journal of Physiology* 160.1, pp. 106–154. eprint: http://jp.physoc.org/content/160/1/106.short.

Jaccard, Paul (1901a). "Distribution de la flore alpine dans le Bassin des Drouces et dans quelques régions voisines". In: *Bulletin de la Société vaudoise des sciences naturelles* 37, pp. 241–272. eprint: http://www.archive.org/details/bulletindelasoc06natugoog.

— (1901b). "Étude comparative de la distribution florale dans une portion des Alpes et des Jura". In: *Bulletin de la Société vaudoise des sciences naturelles* 37, pp. 547–579. eprint: http://www.archive.org/details/bulletindelasoc06natugoog.

— (1907). "La distribution de la flore dans la zone Alpine". In: *Revue générale des Sciences pures et appliquées* 18, pp. 961–967. eprint: http://www.biodiversitylibrary.org/item/41775.

— (1908). "Nouvelles recherches sur la distribution florale". In: *Bulletin de la Société vaudoise des sciences naturelles* 44, pp. 223–270. DOI: http://www.archive.org/details/bulletindelasoc30natugoog.

— (1912). "The Distribution of the Flora in the Alpine Zone". In: *New Phytologist* 11.2, pp. 37–50. ISSN: 0028646X. eprint: http://www.jstor.org/stable/2427226.

Jaiantilal, Abhishek (2012). *MATLAB Random Forest wrapper: MEX interface to Andy Liaw et al.'s C code from R randomForest package (SVN version downloaded 16 Jan 2012)*. eprint: http://code.google.com/p/randomforest-matlab/.

Jain, V. et al. (2007). "Supervised Learning of Image Restoration with Convolutional Networks". In: *IEEE 11th International Conference on Computer Vision (ICCV)*, pp. 1–8. DOI: 10.1109/ICCV.2007.4408909.

Jain, V. et al. (2010). "Boundary Learning by Optimization with Topological Constraints". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2488 –2495. DOI: 10.1109/CVPR.2010.5539950.

Jarrell, Travis A. et al. (2012). "The Connectome of a Decision-Making Neural Network". In: *Science* 337.6093, pp. 437–444. DOI: 10.1126/science.1221762.

Jeong, Won-Ki (2009). "Scalable and Interactive Segmentation and Visualization of Neural Processes in EM Datasets". In: *IEEE Transactions on Visualization and Computer Graph-*

ics 15. Ed. by Johanna Beyer et al., pp. 1505–1514. ISSN: 1077-2626. DOI: 10.1109/TVCG.2009.178.

Jeong, Won-Ki et al. (2010). "SSECRETT and NeuroTrace: Interactive Visualization and Analysis Tools for Large-Scale Neuroscience Datasets". In: *IEEE Computer Graphics and Applications* preprint. eprint: http://people.seas.harvard.edu/~wkjeong/publication/CG_CG&ASI-2009-09-0112.R1_Jeong.pdf.

Jones, Cory et al. (2013). "Semi-automatic neuron segmentation in electron microscopy images via sparse labeling". In: *IEEE 10th International Symposium on Biomedical Imaging (ISBI)*, pp. 1304–1307. DOI: 10.1109/ISBI.2013.6556771.

Jones, J. P. and L. A. Palmer (1987). "The two-dimensional spatial structure of simple receptive fields in cat striate cortex". In: *Journal of Neurophysiology* 58.6, pp. 1187–1211. eprint: http://jn.physiology.org/content/58/6/1187.abstract.

Jurrus, Elizabeth et al. (2008). "An optimal-path approach for neural circuit reconstruction". In: *IEEE 5th International Symposium on Biomedical Imaging: From Nano to Macro (ISBI)*, pp. 1609–1612. DOI: 10.1109/ISBI.2008.4541320.

Jurrus, Elizabeth et al. (2009a). "Axon tracking in serial block-face scanning electron microscopy". In: *Medical Image Analysis* 13.1, pp. 180–188. ISSN: 1361-8415. DOI: 10.1016/j.media.2008.05.002.

Jurrus, Elizabeth et al. (2009b). "Serial Neural Network Classifier for Membrane Detection using a Filter Bank". In: *Fourth international workshop on Microscopic Image Analysis with Applications in Biology (MIAAB)*. eprint: www.sci.utah.edu/~liz/jurrus-miaab2009.pdf.

Jurrus, Elizabeth et al. (2010). "Detection of neuron membranes in electron microscopy images using a serial neural network architecture". In: *Medical Image Analysis* 14.6, pp. 770 –783. ISSN: 1361-8415. DOI: 10.1016/j.media.2010.06.002.

Jurrus, Elizabeth et al. (2013). "Semi-Automated Neuron Boundary Detection and Nonbranching Process Segmentation in Electron Microscopy Images". English. In: *Neuroinformatics* 11.1, pp. 5–29. ISSN: 1539-2791. DOI: 10.1007/s12021-012-9149-y.

Kaplan, E. L. and Paul Meier (1958). "Nonparametric Estimation from Incomplete Observations". English. In: *Journal of the American Statistical Association* 53.282, pp. 457–481. ISSN: 01621459. eprint: http://www.jstor.org/stable/2281868.

Kawrykow, Alexander et al. (2012). "Phylo: A Citizen Science Approach for Improving Multiple Sequence Alignment". In: *PLoS ONE* 7.3, e31362. DOI: 10.1371/journal.pone.0031362.

Kaynig, Verena, Thomas Fuchs, and Joachim Buhmann (2010a). "Geometrical Consistent 3D Tracing of Neuronal Processes in ssTEM Data". In: *Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Ed. by Tianzi Jiang et al. Vol. LNCS 6362. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, pp. 209–216. DOI: 10.1007/978-3-642-15745-5_26.

Kaynig, Verena, Thomas Fuchs, and Joachim M. Buhmann (2010b). "Neuron geometry extraction by perceptual grouping in ssTEM images". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2902–2909. DOI: 10.1109/CVPR.2010.5540029.

Kent, Allen et al. (1955). "Machine literature searching VIII. Operational criteria for designing information retrieval systems". In: *American Documentation* 6.2, pp. 93–101. ISSN: 1936-6108. DOI: 10.1002/asi.5090060209.

Knott, Graham et al. (2008). "Serial Section Scanning Electron Microscopy of Adult Brain Tissue Using Focused Ion Beam Milling". In: *Journal of Neuroscience* 28.12, pp. 2959–2964. DOI: 10.1523/JNEUROSCI.3189-07.2008.

Knowles-Barley, Seymour et al. (2011). "Biologically inspired EM image alignment and neural reconstruction". In: *Bioinformatics* 27.16, pp. 2216–2223. DOI: 10.1093/bioinformatics/btr378.

Koenderink, Jan (1984). "The structure of images". In: *Biological Cybernetics* 50.5, pp. 363–370. ISSN: 0340-1200. DOI: 10.1007/BF00336961.

Köthe, Ullrich (2012). *The VIGRA Computer Vision Library*. eprint: http://hci.iwr.uni-heidelberg.de/vigra/.

Kreshuk, A. et al. (2011). "Automated segmentation of synapses in 3D EM data". In: *IEEE 8th International Symposium on Biomedical Imaging: From Nano to Macro (ISBI)*, pp. 220–223. DOI: 10.1109/ISBI.2011.5872392.

Kulczynski, S. (1927). "Die Pflanzenassoziationen der Pieninen". In: *Bulletin International de l'Académie Polonaise des Sciences et des Lettres. Classe des sciences mathématiques et naturelles. Série B, Sciences naturelles.* 2, pp. 57–203.

Kumar, Ritwik, Amelio Vazquez-Reina, and Hanspeter Pfister (2010). "Radon-Like features and their application to connectomics". In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 186–193. DOI: 10.1109/CVPRW.2010.5543594.

Laptev, Dmitry et al. (2012). "Anisotropic ssTEM Image Segmentation Using Dense Correspondence across Sections". In: *Medical Image Computing and Computer-Assisted Inter-*

vention (MICCAI). Vol. LNCS 7510. Lecture Notes In Computer Science, pp. 323–330. DOI: 10.1007/978-3-642-33415-3_40.

LeCun, Y. (1985). "Une procédure d'apprentissage pour réseau a seuil asymmetrique (a Learning Scheme for Asymmetric Threshold Networks)". In: *Proceedings of Cognitiva* 85, pp. 599–604. eprint: http://yann.lecun.com/exdb/publis/pdf/lecun-85.pdf.

LeCun, Y. et al. (1989). "Backpropagation Applied to Handwritten Zip Code Recognition". In: *Neural Computation* 1.4, pp. 541–551. ISSN: 0899-7667. DOI: 10.1162/neco.1989.1.4.541.

Lesot, M-J., M. Rifqi, and H. Benhadda (2009). "Similarity measures for binary and numerical data: a survey". In: *International Journal of Knowledge Engineering and Soft Data Paradigms* 1.1, pp. 63–84. DOI: 10.1504/IJKESDP.2009.021985.

Liaw, Andy and Matthew Wiener (2002). "Classification and Regression by randomForest". In: *R News* 2.3, pp. 18–22. eprint: http://CRAN.R-project.org/doc/Rnews/.

Lillholm, Martin and Lewis D. Griffin (2008). "Novel image feature alphabets for object recognition". In: *19th International Conference on Pattern Recognition (ICPR)*, pp. 1–4. DOI: 10.1109/ICPR.2008.4761173.

Liu, Ting et al. (2012). "Watershed merge tree classification for electron microscopy image segmentation". In: *21st International Conference on Pattern Recognition (ICPR)*, pp. 133–137. eprint: http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=6460090.

Livet, Jean et al. (2007). "Transgenic strategies for combinatorial expression of fluorescent proteins in the nervous system". In: *Nature* 450.7166, pp. 56–62. ISSN: 0028-0836. DOI: 10.1038/nature06293.

Lowe, David G. (2004). "Distinctive Image Features from Scale-Invariant Keypoints". English. In: *International Journal of Computer Vision* 60.2, pp. 91–110. ISSN: 0920-5691. DOI: 10.1023/B:VISI.0000029664.99615.94.

Lucchi, Aurelien et al. (2010). "A Fully Automated Approach to Segmentation of Irregularly Shaped Cellular Structures in EM Images". In: *Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Ed. by Tianzi Jiang et al. Vol. LNCS 6362. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, pp. 463–471. DOI: 10.1007/978-3-642-15745-5_57.

Lu, Huo, Angelica V. Esquivel, and James M. Bower (2009). "3D electron microscopic reconstruction of segments of rat cerebellar purkinje cell dendrites receiving ascending and

parallel fiber granule cell synaptic inputs". In: *Journal of Comparative Neurology* 514.6, pp. 583–594. ISSN: 1096-9861. DOI: 10.1002/cne.22041.

Macke, Jakob H. et al. (2008). "Contour-propagation algorithms for semi-automated reconstruction of neural processes". In: *Journal of Neuroscience Methods* 167.2, pp. 349–357. ISSN: 0165-0270. DOI: 10.1016/j.jneumeth.2007.07.021.

Maisak, Matthew S. et al. (2013). "A directional tuning map of Drosophila elementary motion detectors". In: *Nature* 500.7461, pp. 212–216. ISSN: 0028-0836. eprint: http://dx.doi.org/10.1038/nature12320.

Markram, Henry (2006). "The Blue Brain Project". In: *Nature Reviews. Neuroscience* 7.2, pp. 153–160. ISSN: 1471-003X. eprint: http://dx.doi.org/10.1038/nrn1848.

McNab, Jennifer A. et al. (2013). "The Human Connectome Project and beyond: Initial applications of 300mT/m gradients". In: *NeuroImage* 80, pp. 234 –245. ISSN: 1053-8119. DOI: http://dx.doi.org/10.1016/j.neuroimage.2013.05.074.

Merchan-Perrez, Angel et al. (2009). "Counting synapses using FIB/SEM microscopy: a true revolution for ultrastructural volume reconstruction". In: *Frontiers in Neuroanatomy* 3, e18. DOI: 10.3389/neuro.05.018.2009.

Mishchenko, Yuriy (2010). "On Optical Detection of Densely Labeled Synapses in Neuropil and Mapping Connectivity with Combinatorially Multiplexed Fluorescent Synaptic Markers". In: *PLoS ONE* 5.1, e8853–. DOI: 10.1371/journal.pone.0008853.

Mishchenko, Yuriy et al. (2010). "Ultrastructural Analysis of Hippocampal Neuropil from the Connectomics Perspective". In: *Neuron* 67.6, pp. 1009–1020. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2010.08.014.

Mohammadi-Gheidari, A., C. W. Hagen, and P. Kruit (2010). "Multibeam scanning electron microscope: Experimental results". In: *Journal of Vacuum Science & Technology B (Proceedings of the 54th international conference on electron, ion and photon beam technology and nanofabrication)*. Vol. 28. 6. AVS, C6G5–C6G10. DOI: 10.1116/1.3498749.

Ochiai, Akira (1957). "Zoogeographic studies on the soleoid fishes found in Japan and its neighbouring regions". In: *Bulletin of the Japanese Society of Scientific Fisheries* 22.9, pp. 526–530. DOI: 10.2331/suisan.22.526.

O'Reilly, Martin T. et al. (2011). "A circle-based method for detection of neural fibre cross-sections in classically stained 2D electron micrographs". In: *Sixth international workshop on Microscopic Image Analysis with Applications in Biology (MIAAB), Heidelberg*. eprint: http://www.miaab.org/miaab-2011-heidelberg-papers.html.

Perkel, Jeffrey M. (2013). In: *Science product article* 339, pp. 350–352. DOI: 10.1126/science.opms.p1300071.

Polak, Mark, Hong Zhang, and Minghong Pi (2009). "An evaluation metric for image segmentation of multiple objects". In: *Image and Vision Computing* 27.8, pp. 1223 –1227. ISSN: 0262-8856. DOI: 10.1016/j.imavis.2008.09.008.

Rand, William M. (1971). "Objective Criteria for the Evaluation of Clustering Methods". In: *Journal of the American Statistical Association* 66.336, pp. 846–850. ISSN: 01621459. DOI: 10.2307/2284239.

Reid, R. Clay and Jose-Manuel Alonso (1995). "Specificity of monosynaptic connections from thalamus to visual cortex". In: *Nature* 378.6554, pp. 281–284. DOI: 10.1038/378281a0.

Rodieck, R.W. (1965). "Quantitative analysis of cat retinal ganglion cell response to visual stimuli". In: *Vision Research* 5.12, pp. 583 –601. ISSN: 0042-6989. DOI: http://dx.doi.org/10.1016/0042-6989(65)90033-7.

Rogers, David J. and Taffee T. Tanimoto (1960). "A Computer Program for Classifying Plants". In: *Science* 132.3434, pp. 1115–1118. ISSN: 00368075. eprint: http://www.jstor.org/stable/1706749.

Rosenblatt, F. (1958). "The perceptron: A probabilistic model for information storage and organization in the brain.[Article]". In: *Psychological Review* 65.6, pp. 386–408.

Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams (1986). "Learning representations by back-propagating errors". In: *Nature* 323.6088, pp. 533–536. eprint: http://dx.doi.org/10.1038/323533a0.

Russel, P.F. and T.R. Rao (1940). "On habitat and association of species of anopheline larvae in south-eastern madras". In: *Journal of Malaria India Institute* 3, pp. 153–178.

Rust, Nicole C. et al. (2006). "How MT cells analyze the motion of visual patterns". In: *Nature Neuroscience* 9.11, pp. 1421–1431. ISSN: 1097-6256. DOI: 10.1038/nn1786.

Sanger, F. et al. (1977). "Nucleotide sequence of bacteriophage φX174 DNA". In: *Nature* 265.5596, pp. 687–695. eprint: http://dx.doi.org/10.1038/265687a0.

Schalek, R. et al. (2012). "ATUM-based SEM for high-speed large-volume biological reconstructions". In: *Microscopy and Microanalysis* 18 (Suppl 2), pp. 572–573. DOI: 10.1017/S1431927612004710.

Schindelin, Johannes et al. (2012). "Fiji: an open-source platform for biological-image analysis". In: *Nature Methodsds* 9.7, pp. 676–682. ISSN: 1548-7091. DOI: 10.1038/nmeth.2019.

Setsompop, K. et al. (2013). "Pushing the limits of in vivo diffusion MRI for the Human Connectome Project". In: *NeuroImage* 80, pp. 220 –233. ISSN: 1053-8119. DOI: http://dx.doi.org/10.1016/j.neuroimage.2013.05.078.

Seung, H. Sebastian (2013). *Eyewire: Play a game to map the brain*. eprint: http://eyewire.org/.

Seyedhosseini, M., M.H. Ellisman, and T. Tasdizen (2013). "Segmentation of mitochondria in electron microscopy images using algebraic curves". In: *IEEE 10th International Symposium on Biomedical Imaging (ISBI)*, pp. 860–863. DOI: 10.1109/ISBI.2013.6556611.

Seyedhosseini, Mojtaba et al. (2011). "Detection of Neuron Membranes in Electron Microscopy Images Using Multi-scale Context and Radon-Like Features". In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2011*. Ed. by Gabor Fichtinger, Anne Martel, and Terry Peters. Vol. 6891. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 670–677. ISBN: 978-3-642-23622-8. DOI: 10.1007/978-3-642-23623-5_84.

Shaar, Nader (2012). *ISBI 2012 Challenge Server: Segmentation of neuronal structures in EM stacks*. eprint: http://brainiac2.mit.edu/isbi_challenge/.

— (2013). *ISBI 2013 Challenge Website: 3D Segmentation of Neurites in EM Images*. eprint: http://brainiac2.mit.edu/SNEMI3D/.

Shroff, Hari et al. (2008). "Live-cell photoactivated localization microscopy of nanoscale adhesion dynamics". In: *Nature Methods* 5.5, pp. 417–423. ISSN: 1548-7091. DOI: 10.1038/nmeth.1202.

Sinsheimer, Robert L. (1989). "The Santa Cruz Workshop–May 1985". In: *Genomics* 5.4, pp. 954–956. ISSN: 0888-7543. DOI: 10.1016/0888-7543(89)90142-0.

Smith, K., A. Carleton, and V. Lepetit (2009). "Fast Ray features for learning irregular shapes". In: *IEEE 12th International Conference on Computer Vision (ICCV)*, pp. 397–404. DOI: 10.1109/ICCV.2009.5459210.

Sohn, Yunkyu et al. (2011). "Topological Cluster Analysis Reveals the Systemic Organization of the <italic>Caenorhabditis elegans</italic> Connectome". In: *PLoS Comput Biol* 7.5, e1001139. DOI: 10.1371/journal.pcbi.1001139.

Sokal, Robert R. and Charles Michener (1958). "A statistical method for evaluating systematic relationships". In: *The university of Kansas Science Bulletin* 38.22, pp. 1409–1438. eprint: http://citebank.org/node/33927.

Sokal, Robert R. and Peter H. Sneath (1963). *Principles of numerical taxonomy*. W.H. Freeman and Company, San Francisco.

Sokal, Robert R. and Peter H. A. Sneath (1973). *Principles of numeric taxonomy*. San Francisco and London: W.H. Freeman and Company.

Sommer, Christoph et al. (2011). "ilastik: Interactive Learning and Segmentation Toolkit". In: *IEEE 8th International Symposium on Biomedical Imaging (ISBI)*. DOI: 10.1109/ISBI.2011.5872394.

Sorenson, T. (1948). "A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons". In: *Det Kongelige Danske videnskabernes selskab Biologiske skrifter* 5, pp. 1–34.

Sporns, Olaf, Giulio Tononi, and Rolf Kotter (2005). "The Human Connectome: A Structural Description of the Human Brain". In: *PLoS Computational Biology* 1.4, e42–. DOI: 10.1371/journal.pcbi.0010042.

Stiles, Joel R. and Thomas M. Bartol (2001). "Chapter 4: Computational Neuroscience: Realistic Modeling for Experimentalists". In: ed. by Erik De Schutter. CRC Press. Chap. Monte Carlo Methods for Simulating Realistic Synaptic Microphysiology Using MCell, pp. 87–127. eprint: http://www.mcell.cnl.salk.edu/Publications/.

Straehle, C. et al. (2011). "Carving: Scalable Interactive Segmentation of Neural Volume Electron Microscopy Images". In: *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Ed. by Gabor Fichtinger, Anne Martel, and Terry Peters. Vol. LNCS 6891. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, pp. 653–660. ISBN: 978-3-642-23622-8. DOI: 10.1007/978-3-642-23623-5_82.

Sulston, J.E., D.G. Albertson, and J.N. Thomson (1980). "The Caenorhabditis elegans male: Postembryonic development of nongonadal structures". In: *Developmental Biology* 78.2, pp. 542 –576. ISSN: 0012-1606. DOI: http://dx.doi.org/10.1016/0012-1606(80)90352-8.

Takemura, Shin-ya et al. (2013). "A visual motion detection circuit suggested by Drosophila connectomics". In: *Nature* 500.7461, pp. 175–181. ISSN: 0028-0836. eprint: http://dx.doi.org/10.1038/nature12450.

Tanaka, K. (1983). "Cross-correlation analysis of geniculostriate neuronal relationships in cats". In: *Journal of Neurophysiology* 49.6, pp. 1303–1318. eprint: http://jn.physiology.org/content/49/6/1303.abstract.

Turaga, Srinivas C. et al. (2009). "Maximin affinity learning of image segmentation". In: *Conference of the Neural Information Processing Systems Foundation (NIPS)*. eprint: http://arxiv.org/abs/0911.5372.

Turaga, Srinivas C. et al. (2010). "Convolutional Networks Can Learn to Generate Affinity Graphs for Image Segmentation". In: *Neural Computation* 22.2, pp. 511–538. ISSN: 0899-7667. DOI: 10.1162/neco.2009.10-08-881.

Tversky, Amos (1977). "Features of Similarity". In: *Psychological Review* 84.4, pp. 327–352. eprint: http://homepage.psy.utexas.edu/homepage/group/loveLAB/love/classes/concepts/Tversky1977.pdf.

Varshney, Lav R. et al. (2011). "Structural Properties of the Caenorhabditis elegans Neuronal Networks". In: *PLoS Computational Biology* 7.2, e1001066. DOI: 10.1371/journal.pcbi.1001066.

Vazquez-Reina, A., E. Miller, and H. Pfister (2009). "Multiphase geometric couplings for the segmentation of neural processes". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2020 –2027. DOI: 10.1109/CVPR.2009.5206524.

Vazquez-Reina, A. et al. (2011). "Segmentation fusion for connectomics". In: *IEEE International Conference on Computer Vision (ICCV)*, pp. 177 –184. DOI: 10.1109/ICCV.2011.6126240.

Veeraraghavan, Ashok et al. (2010). "Increasing Depth Resolution of Electron Microscopy of Neural Circuits using Sparse Tomographic Reconstruction". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1767–1774. DOI: 10.1109/CVPR.2010.5539846.

Venkataraju, Kannan Umadevi et al. (2009). "Automatic markup of neural cell membranes using boosted decision stumps". In: *IEEE 6th International Symposium on Biomedical Imaging (ISBI)*. ISBI'09. Boston, Massachusetts, USA: IEEE Press, pp. 1039–1042. ISBN: 978-1-4244-3931-7. eprint: http://portal.acm.org/citation.cfm?id=1699872.1700138.

Venter, J. Craig et al. (2001). "The Sequence of the Human Genome". In: *Science* 291.5507, pp. 1304–1351. DOI: 10.1126/science.1058040.

Vu, N. and B.S. Manjunath (2008). "Graph cut segmentation of neuronal structures from transmission electron micrographs". In: *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*, pp. 725–728. DOI: 10.1109/ICIP.2008.4711857.

Ward, Samuel et al. (1975). "Electron microscopical reconstruction of the anterior sensory anatomy of the nematode caenorhabditis elegans". In: *The Journal of Comparative Neurology* 160.3, pp. 313–337. ISSN: 1096-9861. DOI: 10.1002/cne.901600305.

Werbos, Paul J. (1974). "Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences". PhD thesis. Harvard University.

White, J. G. et al. (1976). "The Structure of the Ventral Nerve Cord of Caenorhabditis elegans". In: *Philosophical Transactions of the Royal Society of London. B, Biological Sciences* 275.938, pp. 327–348. DOI: 10.1098/rstb.1976.0086.

— (1986). "The Structure of the Nervous System of the Nematode Caenorhabditis elegans". In: *Philosophical Transactions of the Royal Society of London. B, Biological Sciences* 314.1165, pp. 1–340. DOI: 10.1098/rstb.1986.0056.

Wildanger, D. et al. (2009). "A compact STED microscope providing 3D nanoscale resolution". In: *Journal of Microscopy* 236.1, pp. 25–43. DOI: 10.1111/j.1365-2818.2009.03188.x.

Willingham, M. C. and A. V. Rutherford (1984). "The use of osmium-thiocarbohydrazide-osmium (OTO) and ferrocyanide-reduced osmium methods to enhance membrane contrast and preservation in cultured cells". In: *Journal of Histochemistry & Cytochemistry* 32.4, pp. 455–60. DOI: 10.1177/32.4.6323574.

Yang, Huei-Fang and Yoonsuck Choe (2009). "3D volume extraction of densely packed cells in EM data stack by forward and backward graph cuts". In: *IEEE Symposium Computational Intelligence for Multimedia Signal and Vision Processing (CIMSVP)*, pp. 47–52. DOI: 10.1109/CIMSVP.2009.4925647.

Young, Richard A. (1987). "The Gaussian derivative model for spatial vision: I. Retinal mechanisms". In: *Spatial Vision* 2, 273–293(20). DOI: 10.1163/156856887X00222.

Young, Richard A., Ronald M. Lesperance, and W. Weston Meyer (2001). "The Gaussian Derivative model for spatial-temporal vision: I. Cortical model". In: *Spatial Vision* 14 (Issue 3-4), 261–319(58). DOI: 10.1163/156856801753253582.

Yushkevich, Paul A. et al. (2006). "User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability". In: *Neuroimage* 31.3, pp. 1116–1128. ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2006.01.015.

Zhong, Ming and Kenneth R. Hess (2009). *Mean Survival Time from Right Censored Data*. Tech. rep. Working paper 66. Biostats COBRA Preprint series. eprint: http://biostats.bepress.com/cobra/art66.